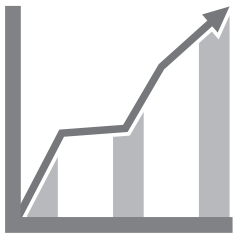


# 从1开始
## 数据分析师成长之路

张旭东　著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

## 内 容 简 介

数据分析行业就像所有新兴行业初期一样，伴随着混乱和盲目，一方面市场上培训机构巧立名目颁发证书，另一方面也有许多国外的著作被生搬硬套过来供自学者学习。本书是第一本结合国内公司实际状况和作者多年数据分析经验，系统而又详尽地介绍数据分析工作的作品。相较于使用 Excel 进行数据统计工作更加专业化、系统化，相较于数据挖掘与编程算法更加易于理解和贴合业务。从简单的制作报表开始和大家一起学习数据分析的五大模块：报表 BI 系统、异常数据分析、解决数据需求、项目性数据分析以及数据建模，为大家全方位、体系化地呈现数据分析到底是什么。



**图书在版编目（CIP）数据**

从 1 开始：数据分析师成长之路 / 张旭东著. —北京：电子工业出版社，2017.1
ISBN 978-7-121-30679-2

Ⅰ. ①从… Ⅱ. ①张… Ⅲ. ①数据处理 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字(2016)第 311433 号

策划编辑：石　倩
责任编辑：石　倩
印　　刷：三河市华成印务有限公司
装　　订：三河市华成印务有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编 100036
开　　本：720×1000　1/16　印张：12.75　字数：204 千字
版　　次：2017 年 1 月第 1 版
印　　次：2017 年 1 月第 1 次印刷
定　　价：49.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：010-51260888-819，faq@phei.com.cn。

# 序言

20 世纪 80 年代，伴随着微型智能计算机的发展，第三次工业革命进入了一个崭新的时代，计算机科学伴随着摩尔定律一路高歌猛进冲进了每个人的生活。从工业化时代转换到互联网时代一个最为突出的特征就是“信息爆炸”，近 30 年来人类生产的信息已超过过去 5000 年信息生产的总和。而当下信息的主要载体是数据库，庞大的信息量对应着庞大的数据量，那么这些承载着庞大信息量的数据处理就显得尤为重要，数据分析作为一门新兴的行业也变得越来越受人瞩目。

就像在计算机行业刚刚火爆的那些年，由于没有现成的体系化的知识，几乎所有人都在摸索中前进。大家的知识一方面来源于相互探讨交流，另一方面借鉴西方发达国家的教材资料。数据分析现在同样没有体系化的知识结构，没有成熟的经验教训，数据分析从业者中一部分是从计算机编程开始做数据挖掘，另一部分是从统计学开始做数据分析，还有一小部分人是凭借着自己的兴趣爱好自己探索着前进。国内对于数据分析的解读一方面偏向于基于 Excel 可视化报表，另一方面偏向于数据挖掘与编程算法，前者太过流于表面，后者又十分晦涩难懂。张旭东的这本《从 1 开始——数据分析师成长之路》算是国内第一本详尽而又系统的介绍数据分析前因后果的书籍了，在保证通俗易懂的同时又有数据分析的深度，作为数据分析的入门书籍的确是相当不错。

数据分析行业一定会伴随着大数据时代的到来逐渐被大家重视和认可，如果你想把握住机会成为大数据时代的弄潮儿，这本书值得一看。

卢斌

中国人民大学高礼研究院执行院长

# 前言

随着大数据这个概念被越来越多的人提起，数据分析与数据挖掘这两个词汇频繁地出现在人们的视野中，越来越得到大家的重视和青睐。从事数据分析工作的这些年，身边不断有人问起数据分析如何入门或是如何做好数据分析，市场也有各类“速成数据分析”或是“零基础数据分析”等培训课程，颇有当年人人都去做产品经理的势头。与此同时在一些问答类网站上出现了许多诸如这样的问题：

“文科生如何转行数据分析？”

“数学基础不好能做数据分析吗？”

“听了某某专家的演讲觉得数据分析很棒，如何入门？”

……

问题下面往往有很多因各种各样的原因推荐的书籍、教程、公众号……内容乏善可陈的同时太容易误导新人，看着着实心痛。

与此同时，通过这些年来的了解和熟悉，身边有太多“盲目”的数据分析从业人员，只是了解了 Excel 中相关图表与统计的功能，在从事分析工作时也有许多的不严谨和漏洞。在一些社区或是平台经常遇到一些人把原始数据直接挂在网上，问该怎么分析数据甚至是通过这些数据能得出什么结论。现在想一想，他们真的适合做数据分析吗？数据保密性的职业素养不说，不经大脑思考地贴数据要

结果的分析员真的能胜任这份工作吗？

写这本书最大的愿望就是能够通过简单的描述让大家对数据分析有一个简单的了解，对自己是否适合这个职位有一个概念，不要盲目从众，能有自己的判断。市场上从零开始入门的教程鱼龙混杂，在入门之前大家首先要考虑这扇门真的适合你吗？

这本书写在数据分析入门之前，会向读者们简单地介绍究竟什么是数据分析，重点放在这个岗位有怎样的要求和特质以及如何才能达到这样的标准，也会简单介绍数据分析岗位未来的职业发展，希望对有志于从事数据分析工作的你有所帮助。

作　者

# 目录

# 第1章 数字、数据、数学

数据作为数据分析的目标与对象无时无刻不充斥在我们的生活中，用数据说话或是数据为王的准则被人们时时刻刻地挂在嘴边，那么数字、数据与数学到底是怎样的存在呢？

## 1.1 数字的起源

从我们的祖先发明结绳计数以来，数字和文字、自然语言就一起作为信息的载体融入生活之中了。数字让人们对日常生活的认识不仅停留在“多”或是“少”，“够”或是“不够”这一简单的逻辑判定上，而是让人们开始有了量化的认识。

在数字还未出现的原始社会，原始人类的生存活动主要停留在觅食中，努力生存下去，“是否”有充足的实物以及“是否”有安全的庇护这些“True or False”的逻辑判断几乎是日常生活的全部内容，人们不需要数字。就好像人们在满足基本生存需求之前不会思考“我从哪里来？要到哪里去？”的哲学问题，原始人类在生存渴望的驱动下面临的问题只是是否有食物？是否有住所？是否存在危险？……

然而随着旧石器时代的到来，祖先开始使用工具来狩猎、种植、生产，这些生产条件的改善使得原始人类的猎物出现了剩余，原先能否生存下去的问题逐渐淡出视线，人们开始思考如何对剩余财产进行储存与分配，数字在这个时候也就应运而生了。数字给了财产以量化的准则和分配的标准，每个人应该分配到多少就有了标准，仓库存储剩余就有了准则，以这样的形式，数字作为承载财产的量化信息出现在人们的视野中。

数字作为一种符号，现存的在使用的就有好多种，诸如阿拉伯数字、罗马数字、中文数字等不同的表现方式，也有诸如二进制、十进制、十六进制等不同的计量方式。但是不管以怎样的形式，数字都是一种符号，如果抛去数字所承载的信息它只能算是一种工具。

只有当数字承载着计量财产信息，计算着天文历法、农忙耕种，估算着投入与产出的比重时，数字才具有价值。

数字不仅是一种符号，数字是一种规范的符号。在我们对数字制定诸如加减乘除这样的数学规则之前数字像文字一样是独立的符号。如果我们用 1、2、3、4 对应吃、穿、住、行，当我们饿了我们就说“1”——饭，当我们冷了我们就需要“2”——穿，当我们需要睡觉就说“3”——住，当我们要走路就说“4”——行（图 1-1）。

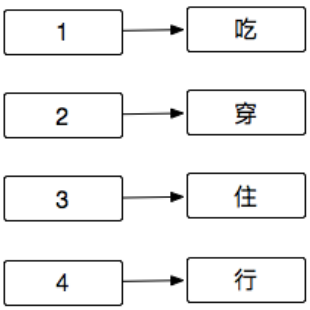



图 1-1

这个时候数字作为一种符号与其他符号相比是不具有特殊性的，但是当我们用数字来量化多少时就把数字赋予了单调性，这里的单调性指的是

$$1<2<3<4<5<6<7\cdots\cdots$$

但是我们不能说

吃>穿>住>行

吃、穿、住、行这样几个符号就不具有单调性。同时我们再赋予数字以加法和减法

$$1+2=3$$

$$4-1=3$$

通过这样的方式，我们的量化符号不再是静态的表示多与少，

它还能表示动态的变化。今年的收成比去年少了，少了多少呢？这就是一个减法的过程，数字对这样一个过程进行了量化和描述，数字的意义就更加丰富了。

## 1.2 数据

什么是数据呢？

数据是对信息一种量化的描述与概括，不仅包含数字所承载的信息，还包含了文字、图片、声音、视频等更多形式所承载的信息，这些都可以统称为数据。严格意义上数据是对客观事物的逻辑归纳，用符号、字母等方式对客观事物进行直观描述。数据是进行各种统计、计算、科学研究或技术设计等所依据的数值，是表达知识的字符的集合。文字、图片、声音、视频等媒介承载的信息可以通过技术手段进行量化，进而转化成数字，这些数字承载了媒介所传达的信息，构成了数据的一部分。

数据作为信息的载体，承载着信息的内容；信息通过数据来表现，让信息变得易识别（图 1-2）。


图 1-2

数字、文字、声音、视频、气味等所有承载的信息的事物都可以理解为数据（图 1-3）。

所以说数据相对于数字是一个更广阔的概念，一切生产活动产生的信息都可以被称之为数据。

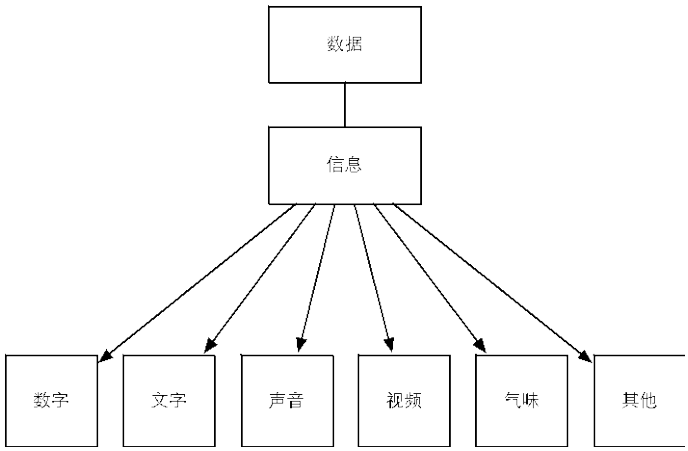



图 1-3

- 按照数据的性质来分，数据可以分为：

a．定位的，如坐标类的数据；

b．定性的，如表示事物属性的数据（居民地、河流、道路等）；

c．定量的，反映事物数量特征的数据，如长度、面积、体积等几何量或重量、速度等物理量；

d．定时的，反映事物时间特性的数据，如年、月、日、时、分、秒等。

- 按照数据的表现形式可以分为：

a．数字数据，如各种统计或量测数据。数字数据在某个区间内是离散的值；

b．模拟数据，由连续函数组成，是指在某个区间连续变化的物理量，又可以分为图形数据（如点、线、面）、符号数据、文字数据和图像数据等，如声音的大小和温度的变化等。

- 按照记录方式又可以分为：

地图、表格、影像、磁带、纸带等。

数据的种类多种多样，记录方式各有不同，按照不同的标准可以分为不同的类别，这些分类的方法主要是帮助我们理解什么是数据及数据能够做什么。我们常常说的大数据时代里的数据来源于此，当我们讨论这些数据时一定要明白它们的来源及意义。

## 1.3 数字与数据

首先数字作为一种符号在其承载信息之后就可以理解为数据，信息作为人类活动传播的内容进行数字化之后变为数据。

我们在接触数学的初期都是从阿拉伯数字 1、2、3、4、5 这些整数开始学起的，之后学习 1/2、1/3、3/5 这些分数，0.1、0.5、3.6 这些小数，进而开始学习-1、-2、-3 这些负数等。我们学习加、减、乘、除的运算法则，开方平方的计算方法，数字配合着运算技巧，这些数字停留在书本上依旧只是数字。就好像常常有人说的：“学数学有什么用？上街买菜用到加减乘除就好了！”这句话里理解的数学就是停留在数字的层面，说的也就是数字的运用。

那么，什么是数据？

请大家读一下下面这句话：

小明今天中午花了 18 元钱吃了一碗牛肉拉面。

你从这句话里能读出什么信息？

如果我把这句话以下面这样的形式做个划分：

小明、今天中午、吃饭、牛肉拉面、18 元

对应如下。

姓名：小明
时间：今天中午（2015/10/20　11：00~14：00）
行为：吃饭

对象：牛肉拉面

单价：18 元

这就可以称之为数据了，这些数据包含了 5 个字段的信息，记录了这样一件事情。

可能会有读者有疑问，上面的文字怎么能算做数据呢？理解起来还是有些别扭。就像我在一开始说的那样，数据不仅包含数字所承载的信息，还包含了文字、图片、声音、视频等更多形式所承载的信息，这些都可以统称为数据。为了给大家一个更直观地认识，这里在原始数据的基础上再加入几条记录：

小明今天中午花了 18 元钱吃了一碗牛肉拉面。

小明今天早上花了 8 元钱吃了一碗阳春面。

小明今天晚上花了 12 元钱吃了一碗刀削面。

小明昨天早上花了 6 元钱吃了一份豆浆油条。

小明昨天中午花了 12 元钱吃了一碗刀削面。

小明昨天晚上花了 10 元钱吃了一碗饺子。

这里统计了小明两天六餐的内容，你能从这六餐里解读出什么信息？

- 小明喜欢吃面，6 餐中 4 餐是面食
- 小明可能是北方人，北方人以面食为主
- 小明应该不是学生，食堂的面条没这么贵

……

这样看来，牛肉拉面、阳春面、刀削面、豆浆油条、饺子这些文本是不是传达出许多信息？所以文本信息作为数据的一种是不是很好理解了。

数字与数据的核心差距就在于前者是向镰刀斧头一样，是我们生活中的工具，而后者是我们生活工作中所有信息的载体。我们使用工具来满足日常生活，更方便的买菜、量化、制定标准，我们使

用信息来描述日常生活，知道过去发生了什么、现在发生了什么、未来将会发生什么……

## 1.4 数学

讨论完数字与数据之后，不妨再来看看数学这个人人敬而远之的学科。许多时候，数据分析入门之所以没那么简单，就在于其门槛相对较高，可以简单地概括为数学门槛和统计学门槛，这两门学科在外行看来注定是枯燥无味的，其实在数据分析的日常中，所使用的数学工具并不多，但是数学与统计学算是底层基础，就好像房子的地基，你看不到，但是必须有。

### 偏见

数学是一门怎样的学科？

数学与语文、英语、物理、化学等学科伴随着我们的学生生涯，数学是我们最早接触的学科之一，其实很多人并不知道为什么要学习数学。从小学到大学漫长的学习生涯中，对数学的认识从加减乘除到微分积分实变复变，长期以来大家对数学这门学科的认识常常是讳莫如深的。有人提出质疑：为什么要学那么多数学？我们上街买菜又用不到。也曾有人在微博上发出投票：数学是否应该滚出高考！有 70%的人投了赞同票。也有部分人尝试为数学这门学科正名，培养逻辑思维等，也算是人云亦云吧。

其实不仅是普罗大众对数学有这样的看法，许多学习数学的数学科学学院的同学也秉持着这样的思想。许多同学都伴随着这样的问题：

我学习数学到底有什么用？

日常生活中哪里能用到？

我毕业之后能干什么？

数学上的许多问题都可以归类为：这也能证明？这也用证明？

在数学上，越是基础的问题越是难，这也是让数科院学生最为困惑的问题。从集合论开始的数学基础涉及的证明题常常让大家崩溃，越来越多的显然的事情需要用数学证明来说明实在让人厌倦和疲惫，一张试卷上只有 6 道令人崩溃的证明题，常常让人翻遍整张试卷而无从下手，许多人在这样一个过程中对数学学习的看法越来越走向及格和毕业。

**例如**

证明：在（0，1）区间中有理数为无数个。（20 分）

上面的题目是不是让你产生这也能证明的想法？细细想来又觉得这也用证明……

凡此种种让数学这门学科包围了越来越多的偏见和错误的认知，“数学无用论”的基调几乎被绝大多数人认同，而在进入数据分析这一行业之前，我们需要来认真地谈一谈数学。

## 数学是什么

维基百科上对数学的定义是这样的：数学是利用符号语言研究数量、结构、变化及空间等概念的一门学科，从某种角度来看属于形式科学的一种。数学透过抽象化和逻辑推理的使用，由计数、计算、量度和对物体形状及运动的观察而产生。数学家们拓展这些概念，为了公式化新的猜想，以及从选定的公理及定义中建立起严谨推导出的定理。

简单来说，数学是一种工具，是一种数字运用的方法。我们生活中的一切无不是由它构建而来的。就好像建筑工人手里的瓦刀或扫地大妈手里的扫帚，作为工具构建我们的世界。同时它也像分子、原子、离子，万事万物无不是由它构建而成，然而我们却看不见它，

感受不到它。数学是基础科学，在此之上建立物理学、工程学，再有桥梁建筑，机械武器，家居日常等。一只简单三脚圆凳，构成了一个圆的内切三角形，与之对应的受力、平衡与结构无不与数学息息相关。当我们用的时候，我们不会在意它，但是它就在那儿。 数学是构建这个世界的最基本工具，穿插在我们的衣食住行之中。

和数学相类似的还有一门学科叫作哲学，身边有很多钻研数学的人对哲学也有很大的兴趣，甚至有人说：世界上只有两个学科，数学和哲学，它们是什么关系？它们构成了整个学术界！也有人说数学是最深刻的哲学，哲学是表象化的数学。其实数学和哲学一样是极其抽象的，数字和符号作为一种表征的方式让数学更容易地进入大家的视野。而抽象的作用是抽取对象中的规律和逻辑，抛弃表现的种种形式，在学习之后能够举一反三，融会贯通，从而触类旁通灵活运用。

一代宗师巴拿赫曾经有这样几句话：

A mathematician is a person who can find analogies between theorems.

习数学者见类比于定理之间。

A better mathematician is one who can see analogies between proofs.

小成者见类比于证明之间。

The best mathematician can notice analogies between theories.

大成者见类比于理论之间。

The ultimate mathematician is one who can see analogies between analogies.

得道高人见类比于类比之间。

简单概括就是数学的四要素：符号、联系、变化、思想。

**数学的第一个要素：符号**

我们常说的 $a$，$b$，$c$，$x$，$y$，$z$，可以代表一个数也可以代表一个集合或是代表一个对象，数学中的各种符号把现实的东西抽象化成符号文字，用来描写和刻画。就好像方程之中的未知数可以任意定名，不影响其参与计算，好似一个实在的数。我们对这个数进行加减乘除，求导积分，需要结果时再把它的实际意义展现出来。这样我们就把实际问题转化为数学问题，又把数学结果应用到实际问题之中解决问题。

**例如：**

小明早上买了 $x$ 个肉包子和 $y$ 个菜包子混装在一个袋子里，问小明第一个吃的包子是肉包子的可能性有多大？如果已知小明第一个吃到的是肉包子之后第二个吃到的是菜包子的可能性有多大？

答：我们用 A 代表肉包子，B 代表菜包子，第一个吃到肉包子的可能性就是 P(A)，已知小明第一个吃到的是肉包子之后第二个吃到的是菜包子的可能性就是 P(B|A)。

**数学的第二个要素：联系**

数学的科技树在漫长的发展过程中从简单的集合论开始发展到拓扑学、图论等，在这样一个过程中我们可能接触了解析几何、线性代数、数学分析。几何专注于形状，代数专注于结构，分析专注于变化，而这三者又是相互联系相互转化的。几何用坐标标明空间与位置，进而转化成向量和矩阵，需要用代数的手段来解决问题，与此同时，当数量不断变化时又需要有极限思想，进而联系到数学分析。就好像计算机中精通了一门编程语言就能触类旁通很快上手其他语言。数学的各个科技树相互联系又各有千秋，并向不同的方向不断前进。

**例如：**

地图上 A 点的坐标是$(x,y,z)$，B 点的坐标是$(a,b,c)$，求 A 到 B 之间的距离。

答：距离 D=$\sqrt{(x-a)^2+(y-b)^2+(z-c)^2}$

**数学的第三个要素：变化**

变化的根本可以理解为 1 即 2，2 即无穷。

当我们证明了一件事情在第 1 维度是可行的时候，就会把它拓展到 2 个维度上是否可行。当我们发现它在 2 个维度上可行的时候我们再想它在 3 个维度、4 个维度、$n$ 个维度上是否同样有效。不知大家是否记得我们学习方程的过程：从一元一次方程到二元一次方程组，再到三元、四元、$N$ 元方程组。

一元一次方程：4×X+2=10

二元一次方程组：

$$2\times X+3\times Y=10$$

$$4\times X+5\times Y=18$$

三元一次方程组：

$$2\times X+3\times Y+4\times Z=14$$

$$4\times X+5\times Y+6\times Z=24$$

$$6\times X+7\times Y+8\times Z=34$$

$N$ 元一次方程组：

$$A\times X=B$$

其中：A 为 $n\times n$ 数字矩阵，X 为 $x_1$，$x_2$…$x_n$ 矩阵，B 为 $n\times 1$ 数字矩阵。

在这样一个对方程从熟悉到了解再到掌握的过程中，也即 1 到 $n$ 过程，在方程的种种变化之中，解方程的手法不变，或者说是思想是始终如一的。

**数学的第四个要素：思想**

数学的思想从美索不达米亚的符号学开始，再到埃及数学、古希腊数学、欧几里得和阿波罗尼斯、亚历山大时期、印度数学和阿拉伯数字、中世纪欧洲的数学，一直到文艺复兴再到微积分的创立，数学的思想汇集了不同方面的成果与点滴积累而成。我们需要知道，数学常常需要几十年甚至几百年的努力才能迈出有意义的一步。数学这门学科的本身并没有锤炼得天衣无缝，即使是那些已经取得的成就，也常常只是一个开始，许多缺陷有待补充，或许真正的扩展还有待创造。数学思想经历了世世代代的汇聚与积累，数学的历史积淀与成长汇聚成了今天我们所谓的数学思想，数学的底蕴深厚，思想源远流长！

数学并不是大家想象得那样高不可攀，遥不可及，数学是一种工具，它充斥在我们生活的每一个角落。它就像构成物质的原子无处不在而有时候又无迹可寻，接受它并以平和的心态理解它，在你需要的时候想办法驾驭它，你会发现数学比你想象中要简单。

其实，如果我们把数学当作摄影、绘画、音乐、舞蹈来学，不要望而生畏，那么数学只不过是一种十分古老的技术甚至说可以是艺术。它博大精深而又晦涩难懂的外表，容易让人敬畏和害怕，但其实数学很有趣的（People don't think math is easy, just because they don't know how complicated life is）。

## 1.5 统计学

兴趣是最好的老师，但是对数学感兴趣着实不是一件简单的事情。相较于数学，统计学在日常生活中的应用要明显而又简单得多。

我们日常生活中接触的求和、平均值、中位数等其实都算是统计学的一部分，统计学有一个非常经典的理论叫作回归分析，我想接触过高等数学的你们一定听说过。当然，随着时间的推移你可能已经忘记了回归分析是怎样一个概念，接下来我们用一个相对趣味的方法给大家介绍下回归分析吧！

从前有个好奇的小明在自家的院子里种豆子，闲来无事他就一个人默默地数豆子，一个、两个、三个、四个……他发现有的豆子个头比较大，有的豆子个头比较小，好奇的小明不禁想要问：同一批次种下的豆子上为什么会结出不同大小的豆子？通过一株一株地看自己种的豆子，他做了一个猜想：大豆子的后代会比较小，小豆子的后代会比较大，为了验证这一猜想他找了全国各地的小伙伴帮忙从相同的袋子里面取不同种类的豆子回去种，然后豆子在他们的悉心照料下渐渐长大，这个好奇的人就把豆子全都要了回来，一个一个标记大小，然后发现自己真聪明啊，之前的假设是正确的，大豆子的后代会逐渐变小，小豆子的后代会逐渐变大，也就是说大家的个头趋于向平均值附近靠拢，他把这种现象取名：返祖。后来大家发现这种返祖的现象出现在了许多植物上，人们对此开始进行学术上的研究，也因此逐渐把返祖更改为学术的回归。

这样看来，回归分析里的回归是不是就简单形象了许多呢？那么我们再进一步地来说一说回归里量化的一些概念。

平均值就是用来衡量回归标准的一个方法，数据围绕着这个平均值波动，并且有向平均值靠拢的趋势即为回归。如图 1-4 所示，我们看到一条曲线围绕着中间的一条直线在上下波动，从某种意义上说我们可以把这条直线理解为这条曲线的回归线，平均值的思想在某种程度上也来源于此。

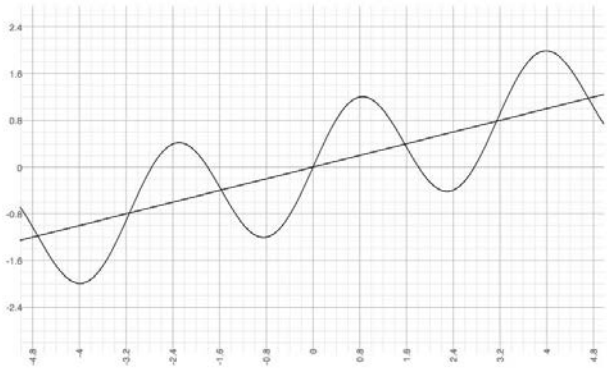


图 1-4

相比较上一张回归曲线，我们不妨看一看图 1-5。

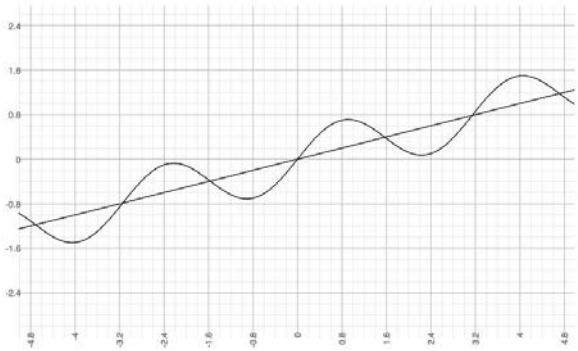


图 1-5

显而易见，图 1-4 和图 1-5 的一个显著不同就是波峰和波谷距离平均线的距离一大一小，在数学上我们用方差来解释这一差异。

$$s^2 = \frac{1}{n}[(x_1 - x)^2 + (x_2 - x)^2 + \cdots + (x_n - x)^2]$$

其中 $x$ 为需要回归的均值，可以理解为上面两张图中的直线，$x_1$~$x_n$ 为各个观测点的数值大小，$n$ 为变量的个数，这样方差就是各个变量与平均值的差值的平方的平均值。因为差值存在正值和负值，

所以平方在这里的作用就是把一切差异平方为正数，避免正负之和相互削弱。这样我们就了解和认识了方差，当我们翻开统计学的课本，书本上一个又一个原先让你头疼的定义和定理其实都是为了解决生活中遇到的问题。

统计学是通过搜索、整理、分析、描述数据等手段，以达到推断所测对象的本质，甚至预测对象未来的一门综合性科学。统计学用到了大量的数学及其他学科的专业知识，其应用范围几乎覆盖了社会科学和自然科学的各个领域。目前国内许多高校都开设了统计学课程，大多是以数学为基础延伸到统计学的分支。我们在进行统计学学习时有许多成熟的理论和体系化的思想，而这些思想与理论常常伴随着专业的术语和难以理解的逻辑，需要将统计学生活化去理解它们。

有一次，突然想到一个问题：在抛硬币事件中，如果在抛了足够多次数后，正面出现的次数远大于反面出现的次数，我们的概率论会变成什么样？世界会因此而改变吗？后来就这个问题向教授《随机过程》的老师请教了一下，略有所想。

在一开始老师是直接否定这一假设的，他认为这是不可能的，因为硬币的均匀问题，以及外界干扰都是不可避免的。但是我坚持用一个数学上常用的“理想状况下”理论，让老师对这一猜想做一个解释与判断。对于我这样“偏执”的学生老师也比较无奈吧，他接受了我的假设，但他说这只会影响客观概率学，不会影响主观概率学，并且主观概率学是概率论的主要内容。即使这样，这个假设对客观概率学的影响不会很大，就好像牛顿的运动定律，即使在未来被认为是存在错误的，但是我们日常生活中仍然使用它，爱因斯坦的相对论是一个更完美的体系，但是相对论包容了牛顿力学而没有排斥它。如果未来出现如我刚才所说的假设，我们会有一个更完美的体系去解释它并且能够包容现在的理论体系。不仅如此，在人类科学发展的几百年来，现在的理论体系经历了无数的实验和实践检验，是绝对可靠的。

首先，我对于老师前面的说法表示完全接受和理解，但是对于后半部分我反驳道：我们的很多理论都是建立在公理之上，这是比较危险的，历史上有很多当时人人都认为对的公理到头来却是错误的。不仅如此，人类科学发展的历程从整个宇宙的起源来看实在是太短暂了，这短短的数百年，何以检验宇宙间最基本的规律？

老师对于我这种想法直接回应的是：不可知论，这种想法是很危险的。现在国外很多书籍宣传一些不够合理的东西，但是由于国外的出版自由很多东西是会误导人的，你们别被这些东西误导。当我表示我看的是霍金的书籍时，老师认为我们无法理解那些太深的东西。然后由于时间关系，在我看来本次谈话就不欢而散了，对于我的质疑老师就是给了我一个“不可知论”的标签。

对于年长的老师我一向是敬畏的，他们的思想可能比较保守，对于这些类似于“空想主义”与“虚无主义”予以排斥。只是年轻人不分是非，一味地接受一些新奇的东西或许是危险的吧！并且老师认为这些与众不同的东西包含有哗众取宠和虚而不实的成分，或许我就是那个哗众取宠者。

总的来说，数学作为一门学科并没有大家想象中那么复杂，如果我们能够以理解小说、电影、时事政治的角度来理解数学，会发现数学很简单。

第2章

# 分析、逻辑与思维

作为一名数据分析师，最重要的当然是分析能力，然而就好像有些人偏向理性思维，有些人偏向感性思维，当我们遇到一个问题时不同的思考方式肯定会导致不同的结果。那么作为数据分析师需要一种怎样的思维方式才能满足要求呢？又需要怎样的分析能力才能让自己的工作出彩呢？

## 2.1 描述、概括、分析

大家在日常生活中经常会听到这些词汇：描述、概括、分析、知道、认识、了解、熟悉、掌握等。这些似乎意思差不多的词汇，粗略看起来并没有什么区别，但是许多时候是说者有心而听者无意，数据分析尤其如此。我们需要描述一个事件还是分析一个事件？这两者中间大有区别，为了便于大家理解先来讲个故事吧。

慵懒的下午，你坐在咖啡馆里看窗外人来人往，这时突然有一位美女闯入了你的眼帘，惊艳了时光，叨扰了岁月。在你的注视中美女就那么徐徐地走了，而你仍旧久久不能忘怀，难得这样的心动时刻，你需要把它记录下来：

2015 年 10 月 21 日，星期三，天气如同心情一样好，邂逅一美女，撰文以记之。

她就那么突然地闯入我的视线，像一只骄傲的猫，带着比肩的短发，蚕眉冷艳，眼波流转；鼻梁不高但棱角分明，唇不红艳自带一份雅致；黑色的小皮鞋轻快地敲打着地砖，颀秀的两条腿包裹在粉色的丝袜中傲娇而不媚俗；白色毛衣披风就那么搭在肩上欲滑将落……

她就那么徐徐地走着，带着独特地隐藏在优雅中的俏皮，伴随着一丝倔强和傲气，轻快又不显急躁地走着……

矫健的步伐配合着摇曳的臂摆透漏了内心的快乐与活力，让人不禁想象这个女孩不管在工作中还是生活中应该都是乐观的吧，平时应该比较爱笑，周围朋友也会很多，应该会很好相处吧！我能不能成为她的朋

友呢?

亲爱的朋友，能不能从上面的一段矫情的日记里面说出哪里是“描述”哪里是“概括”哪里是“分析”呢?

## 描述

抽象来说，描述就是对事物或是对象的直接描写，就好像上文中这个姑娘眼睛、鼻子、嘴唇长什么样，这是对这个对象的客观印象,就好像画画时选择的颜色,我选择红色颜料来描绘他的嘴唇。如果我们把描述这样一个概念对应到数据上可以理解为这一堆数据“长什么样”，按照这样一个标准我们尝试着描述一堆数据。通过对数据的描述能够让别人通过这些描述的话语感受到数据的真实面貌。

对于对人体外貌的描述再详细生动都不如直接看到被描述的这个人，或者给这个人拍一张照片也能直观地反映其外貌。而对于数据来说，直接看数据可能什么都看不出来，而通过对数据的描述反而能让我们更加清晰地看到数据真实的面貌。在了解此间差异之前我们不妨先熟悉几个描述性的统计变量：平均数、众数、中位数、方差、极差、四分位点，这些指标就好像一堆数据的“鼻子”、“眼睛”、“嘴唇”。平均数不用介绍大家都知道，下面介绍下其他几个数据指标：

- 众数：数据中出现频率最高的数值，比如“面条”就可以算做小明数据中的众数。
- 中位数：将数据从小到大排列，位置处于中间的数值。
- 方差：每个数据与平均值的差值的平方，再取平均值。
- 极差：最大数减去最小数。
- 上/下四分位点：将数据从大到小排列，位置处于前 1/4 或是后 1/4 的数值。

**例如**

下面数据记录了小明参加射箭俱乐部时击中的环数：

1　1　2　2　3　5　5　5　6　7　7

上述数据的各项指标如下：

- 平均数=44/11=4
- 众数=5（5 出现 3 次）
- 中位数=5
- 方差=4
- 极差=7-1=6
- 上四分位点=6
- 下四分位点=2

我们一般会用上述的 6 个指标来描述一组数据的“长相”，平均值用来展示整体的平均水平，众数用来展示数据点主要集中的范围，中位数用来与平均数进行对比判断数据是否平滑，方差用来判断数据波动情况。

到这里，我们发现通过对一组数据的平均数、众数、中位数、方差、极差、四分位点进行解读，很容易对这一批数字有具体的认识，而直接看数字可能就感受不到这些信息。不仅如此，我们在数学统计的过程中常常面临着成千上万的数字，如果把这些数字全部罗列在屏幕上可能很难看出什么名堂来，而通过上述 6 个指标能让这些庞大繁杂的数据一目了然，虽不见数据却也知道数据长什么样，这就是描述性统计变量。

## 概括

数据上的概括是形成概念的一种过程，可以理解为基于历史的经验，把大脑中所描述的对象中某些的特征特质抽离出来并形成一种认识，就好像上文中对女孩气质的概括。气质是基于这个女孩走路的姿势、穿衣的风格及面部表情等元素综合在一起，然后基于历

史对“气质”这样一件事情的概念得出的结论。气质是不可以依靠眼睛感受的自然光线来直接获取的，而是需要收集这个人的所有细节描写的信息，形成对这个人的整体印象，然后从整体印象中抽离出“气质”这个充满概括属性的说法。

如果将概括这样的概念引入到数据分析中，最常见的就是正太分布、均匀分布等。为了给大家一个直观的印象，均匀分布可以理解为掷一枚均匀的骰子，各个点数出现的概率是均等的，每次实验都把这些点数记录下来并计算它们出现的概率，每个数字出现的概率就服从均匀分布。

假设我们抛了 10000 次均匀的骰子，记录每一次的点数，会得到这样一组数据：

1、5、2、4、3、2、6……4、6、1

计算 1~6 出现的概率，假设 X 表示点数，P 表示概率，会发现：

P（X=1）≈1/6

P（X=2）≈1/6

P（X=3）≈1/6

P（X=4）≈1/6

P（X=5）≈1/6

P（X=6）≈1/6

于是我们说点数 X 服从离散分布，如图 2-1 所示。

同样的正太分布可以理解为大家都趋向于中间一点的分布。假设我们测量 A 地区 1000 名 20 岁男生的身高，我们会得到这样一组数据：

1.70、1.75、1.82、1.75、1.76……1.81、1.75、1.69、1.78

我们统计每一个身高段的人数如表 2-1 所示。

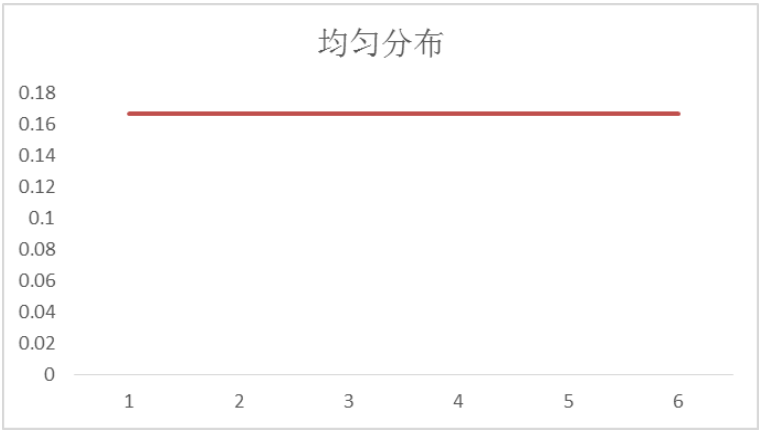



图 2-1

表 2-1

| 组　　别 | 人　　数 |
|---|---|
| A 组 | 20 |
| B 组 | 52 |
| C 组 | 127 |
| D 组 | 317 |
| E 组 | 284 |
| F 组 | 126 |
| G 组 | 58 |
| H 组 | 16 |

画一个柱状图，如图 2-2 所示。

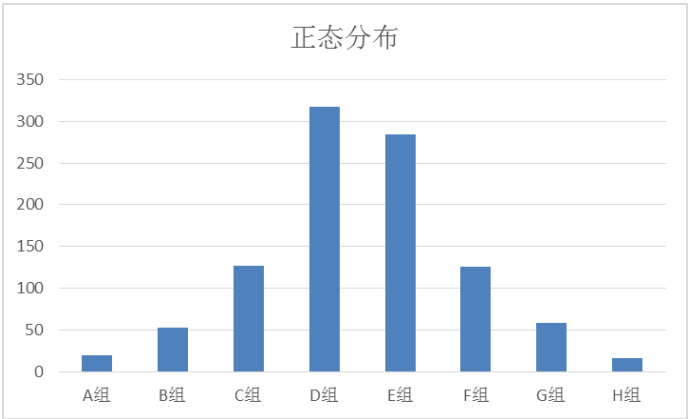



图 2-2

我们把这类数据称之为服从正态分布。

概括的意义在于用一两个简单的概念就能传达出大量的信息，就好像你说某某姑娘“御姐范”、“女王范”、“萝莉范”，我说这个数据服从正态分布、均匀分布、泊松分布。从数据的描述性变量中抽取关键元素，结合已经掌握的经验知识给予数据一个概括：均值为0，方差为 1 的正态分布数据，同业人员听完就基本了解这组数据的特征了。所以说概括是在具象描述的基础上抽离出的概念与总结，结合之前的描述性统计的掌握，当我们面对大量数据时我们先进行描述性统计，然后在基于此给出一个概括就把这样一组庞大的数据信息传递出来了。

到这里，基本可以看到描述与概括的意义了，在庞大繁杂的数据中我们需要一些东西来了解数据，掌握数据的特点，知悉数据的结构，才能为下一步的分析做准备。

## 分析

抽象来说，分析是将研究对象的整体分为各个部分、方面、因素和层次，并分别加以考察的认识活动，也可以通俗地解释为发现隐藏在细节中的“魔鬼”。分析的有效性建立在这样一个共识之上：

一切结果都是有原因的。就好像有些人说的：没有无缘无故的好，也没有无缘无故的坏。

上文中通过对姑娘穿着、打扮、步态等细节的分析，就能分析出这个姑娘的性格与心态，甚至能推断出她的职业、职位等信息，这些都是通过一些外在的可见的细节，通过描述获取细节，再通过概括抽象总结，以此为基础进行抽丝剥茧得到想要的结论。分析区别于描述与概括一个非常重要的特征是以结果为导向，而结果为导向的前提是目的很明确。比如上文中我的目的是了解这个女孩的生活工作习惯，为了达到这样一个目标，我基于她的穿着打扮等信息做出分析，给出结论。同样，当我们面对数据做分析时也一定是以目标为前提，以结果为导向的。

假设我们又抽取了 B 地 1000 名 20 岁男性，测量他们的身高，得到一组数据：

1.69、1.77、1.81、1.74、1.76……1.80、1.74、1.68、1.75

我们把 A 地的 1000 条数据和 B 地的 1000 条数据放在一起组成了有 2000 个观测值的矩阵，我想要知道 A 地男生的身高与 B 地男生的身高的差异情况，该怎么分析呢？

比较描述性统计变量：

均值 $\mu1= \mu2$

方差 $\sigma1= \sigma2$

上下四分位点：QU1=QU2、QL1=QL2

比较数据分布

正态分布的均值与方差

进阶分析

做 T-test 检验

……

我们看到对数据的描述与概括在分析中起了作用，同时还有单独的统计方法 T-Test，如果描述与概括是向别人呈现一组数据，那么分析就是从描述与概括中抽离出能够实现目标的元素，拆分直接实现目的得到结论：总体上 A 地 20 岁男生的身高要高于 B 地男生。

对于一名数据分析人员来说，分析数据的目的性尤为重要，这个会在之后的章节着重说明，此处不做赘述。

总的来说：

- 描述的意义在于让别人知道这个人的外形，这个数据集的长相。
- 概括的意义在于从整体上对对象有一个进一步的了解和认识。
- 分析的特点在于为了达成一个目标而对对象进行一步步地探索和挖掘。

## 2.2 逻辑思维

对于逻辑思维这样一个概念，大家并不陌生，大家对逻辑思维的主观印象是理性、客观、公正等。而之所以能够产生这样印象的思维其实是感性思维，对一件事物的直观感受我们称为感性思维。但是如果能够十分细致地说明为什么逻辑思维是一种理性的思维，甚至罗列出各种原因，在此之后你再说出逻辑思维是一种理性的思维这样的判断，那么这个判断就是理性的了。

逻辑思维往往伴随着理性的思考与决策，而感性思维往往伴随着情绪与冲动。给大家分享一个小故事：

华尔街一家赫赫有名的证券交易公司 A，在“9.11”事件之后凭借幸存下来的资源疯狂地进行资本积累，短时间内在华尔街混得风生水起，然而树大招风，监管机构也开始盯上了这家证券公司。CEO 意识到公司内部的确存在内幕交易的问题，同时大家也都随性而为惯了，短时间内很难上纲上线规范制度，于是 CEO 准备杀鸡儆猴。在一次模拟 FBI

突击检查之后抓住了一个典型员工 K，当场解雇了该员工。K 非常愤怒，扬言要报复该 CEO。

至此，我们的小故事告一段落，感性地来说这个故事讲得很一般，不够生动形象也不够娓娓动听引人入胜，但是我在表达这一件事情的过程中清晰地交代了：

时间：“9.11”之后

地点：华尔街

人物：CEO、K、FBI

事件：杀鸡儆猴

许多时候我们说话时都会凭借本能来表达，有些人的语言天分比较好能说会道，我们说他情商比较高。有些人说话前言不搭后语、思路混乱、言不达意，我们说他情商低。事实上说话的逻辑隐藏在这些“能说会道”与“情商高”之后，这些作为“天赋”的东西之所以可以被培养和练习都源于其背后的逻辑，只是一个擅长表达的人忽略了他的逻辑，这种逻辑贯穿在他说话的始终，就像呼吸空气一样无法察觉。讲述者在表达的过程中自然而然地讲清楚了：时间、地点、人物、事件，这就是逻辑。

我们继续说故事：

这个恼羞成怒的员工想要报复 CEO，非常好理解，一个公司有那么多人，你要杀鸡儆猴为什么拿我开刀，当着那么多人的面炒了我的鱿鱼，我要报复你！为了防止该员工产生过激行为，CEO 派出了心理学家 M 帮忙解决这个问题。M 找到 K，发生了如下谈话。

M：记得你上次这么愤怒的时候是什么时候吗？

K：小学的时候一个老师当众骂我：一辈子都不会有什么出息！

M：你后来怎么做的？

K：当我赚了人生的第一个一百万的时候，我写信狠狠地羞辱了他一番！

M：你觉得有快感吗？

K：有！

M：这种快感持续了多久？

K：（沉思）一天吧！

M：后来呢？

K：日子还是照常进行。

M：我知道这次 CEO 当众解雇你的事情让你很愤怒，你想报复他我很理解，我专程来代表他向你道歉。其实你有两种选择：一种是疯狂地报复他，这很容易，现在媒体和检察院需要这些信息，很容易引起轰动的效果。但是问题是，以 CEO 在华尔街的影响力他会在整个金融行业封杀你，你的储蓄够你无忧无虑地度过下半生吗？

M：另一种方案是接受我的道歉，CEO 会推荐你去另外一家薪资不低于 A 公司的企业，并且 CEO 欠你一个人情。

K：你是在威胁我吗？

M：不，我只是希望你能够做出理性的选择！

小故事到此结束。

心理学家 M 把这个问题解决得很漂亮，足够理性，但他的逻辑是什么呢？仅仅是陈述事实和利害关系吗？

K 的想法很简单：不爽→报复。

M 的逻辑：先回顾历史同类事件拉近距离打开话题，然后总结历史同类事件的感受和经验，找到现在问题与历史问题的相似点，最后剖析利害关系让对方自己做选择。

大家看到 M 在对话的全程都是在引导 K 聊天，M 在发问，K 自己陈述利害关系。最后通过 M 的引导 K 自己选择了 M 希望 K 做的事情。

通过这个小故事大家能感受到语言表达之后隐藏的逻辑，帮助你“思路清晰”，让你“善于表达”，同时也可以在高手是“轻轻松

松”化险为夷之后思考他是如何做到的。

逻辑思维是一个及其抽象的概念，培养逻辑思维需要在日常生活中不断进行理性思考，发现事情背后的真相。许多时候对生活中一些小的事情和问题保持好奇心和勤思考的能力可以锻炼自己的逻辑思维。

犹记得还在上大学的时候，我和小 J 一起在学校西门买了七个包子，我四个，他三个。我的包子为两个豆腐的，两个肉的。之后的争辩是围绕这四个包子展开的，为了不让我们的读者混乱，我们不妨设我的包子为：豆腐 A，豆腐 B，肉包 A，肉包 B；好，下面进入正题。我们在草坪旁吃包子，我突然发现，我先吃的两个包子都是肉的，突发感慨：“今天运气不错啊，连续吃两个肉的！嘿！小 J，连续吃到两个肉包子这个事件的概率是多少？”

小 J 推推眼镜说道：“这个问题简单，可以采用捆绑法，把包子 A 和包子 B 捆绑在一起，四个包子分为三类，事件 A 的出现的次数为 C（3，1），P（A）的值 1/3。”

我说：“不对，豆腐 A 和豆腐 B 在吃的过程中不加区分，包子 A=包子 B，故应该把概率乘以 2！P=2/3”。

“恩，好像有道理！不对，不对，豆腐包与肉包为等价事件，那么连续吃到两个豆腐的概率也是 2/3，但是 2/3+2/3=4/3，违背了 P≤1 的原则，是错误的！”小丁补充道。

“恩，是的，那么我们直接分开讨论吧。事件域的总体为 A（4，4），由于包子不加区分性，再除以 C（2，1）× C（2，1），这样有六种情况，又吃到肉包子为 C（4，2），接下来再进行排序，采用插空法，把这两个包子插进去，共有 C3（1）种插法，有点复杂……”

到这里，我想大家一定开始晕了吧，我们当时也很晕。不过接下来我们发挥无所畏惧的革命精神，采用了挡板法、替代法、虚位以待法……

搞了好久，我们终于发现答案是 1/6，再次重申事件 A 是：我直接连续吃到两个肉包子。最后，我惊奇地发现，答案就是开始时吃到肉包的概率为 1/2，之后为 1/3，P=1/2×1/3。数学真是无处不在。

为了体现学数学的孩子的钻研精神，我们决定讨论肉包与豆腐包是否为独立事件，它们的相关系数、协方差……我们甚至还考虑了老板娘装包子时是否为一起拿的，这样就影响了事件的独立性……最后的最后，我们失望地发现，由于老板娘是把同类包子放在一起的，所以我要么连续吃到两个肉的，要么连续吃到两个豆腐的，好失望……

另一次，我与小 J 一起骑车盘旋而上学校的东吴桥，他走外圈，我走内圈，这时候我们遇到一个问题：谁上坡时比较费力？（指的是坡度比较大，或者说路面仰角比较大。）

小 J 不假思索地说：“一样大！”

我在想我们从同一起点出发到达同一高度，我走的路程比他走的路程明显短一大截。就好比我用一米的路程爬上去，他用十米的路程爬上去，那么谁更费力？谁的坡度大？（注：这是一个单调递增的过程。可以类比到谁的斜率更大？）

但是，如果我们换个思路，把环形路面变成阶梯状这个问题就方便思考了，显然在单位距离内内圈的阶梯数要明显多于外圈（这个旋转楼梯见过的都知道内圈明显比外圈密集），阶梯数多，显然高度差就大，可以说是内圈的坡度比较大。但是又引申出一个问题，不妨设单位距离为 X，当 X 趋近于 0 怎么办？当 X 无限小时，可以说是那一点的斜率是怎么样的呢？是否内圈与外圈的斜率是一样的呢？毕竟路面是连续的没有间断点。

最后我们讨论可以把路面分成 $n$ 个阶梯，不妨让 $n$ 趋近正无穷，那么无论在多小的 X 内都有 $n$ 个阶梯的存在，自然而然回到那个解题的斜率问题了……

生活中这样的案例仍旧很多，无聊地打发时光的日子里我们的逻辑思维得到锻炼，许多问题都因为争执不下而不了了之，但是这不妨碍我们在这样的过程中体验思考的乐趣。

# 第3章 大数据到底是什么

Big data is like teenage sex:

Everyone talks about it,

Nobody really knows how to do it,

Everyone thinks everyone else is doing it,

So everyone claims they are doing it...

## 3.1 时代的现状

伴随着互联网的急速扩张，大数据和云计算、O2O、物联网、互联网+等一系列概念充斥着我们的眼球，无论是涂子沛老师的《大数据》还是舍恩·伯格的《大数据时代》都向我们展示了一种新的可能，关于大数据时代的种种预想和猜测众说纷纭，用数据说话一时间成为当下的潮流。无论是否读过上面说的两本书，你真的了解大数据是什么吗?

这里引入业内一个经典的嘲讽：

Big data is like teenage sex:

Everyone talks about it,

Nobody really knows how to do it,

Everyone thinks everyone else is doing it,

So everyone claims they are doing it...

大数据这个概念以这样的形象被大家所熟知，有好的一面也有坏的一面。好的方面是大家开始逐渐有这样一个概念，并且开始重视起来，只有大家都重视了大数据才可能在未来成为一种趋势，甚至是开创一个大数据时代。而坏的方面则是大数据被人们过度解读，甚至是胡乱解读。移动互联网时代的风口把一切都吹在了天上，大

数据也成为一种潮流与标签，张口闭口都会带着这个的词汇，难免会让人产生反感。

我们不妨换一个角度思考，就好像电商刚刚出现的那段时间，网上商品良莠不齐、交易安全众说纷纭。电商作为一个全新的概念冲进人们的生活有人抵触也有人吹捧，看好者觉得他会取代一切，看空者徘徊不前，不敢前进。然而仅仅经过几年时间的发展，电商已俨然成为一种消费方式融入我们的生活。没有摧枯拉朽的导致实体店的倒闭，也没有因为安全或是造假问题而停滞不前。大数据就是出于这样一个关口，有人过度吹捧也有人不屑一顾，但是时代的发展终归会让一切趋于平静，你信或是不信，大数据都会来的。

其实，当我们环顾周围正在发生的变化，我们看到了数据公司正像雨后春笋一样兴起，各类数据供应商无论是通过线上 API 接口还是线下数据修复都能对外提供包含身份数据、电商数据、新闻数据、社交数据、征信数据等。在这样的一个过程中，大数据崛起最明显的一个行业就是在征信领域。由于移动互联网的蓬勃发展，网络公司通过一个 APP 或是一个 Web 页面都能从用户手中获得庞大的数据量，而这些数据除了满足日常的业务运营需要就那么静静地躺在那里。而这时，随着互联网金融的蓬勃发展，大家发现由于中国征信体制不健全的原因。人们需要越来越多的数据来对一个用户进行风险把控和信用评估，这时大家就开始了对信用领域数据的探索。

最先被人们认可的数据是身份证信息验证，因为这个数据从公安部对外提供开始就一直被人们用来验证一个人与身份信息的匹配度真实性。接下来是学籍认证伴随着大学生市场的开放逐步进入人们的视野。但是，由于网民群体的反欺诈意识较为薄弱，越来越多的身份资料被人伪造和冒用。于是，人们引用了手机运营商数据：要求用户提供运营商服务密码，这样就能抓取用户过去六个月的通话详单。在之后，随着科技的进步和 SDK 技术的成熟，大家开始抓取用户设备信息。

渐渐地，人们开始不仅局限于用户的个体属性类数据，而开始

研究一个人的新闻阅读习惯是否会影响用户的信用情况，接下来步入人们视野的是电商消费数据、GPS 信息、水电煤使用数据、APP 使用数据……

现在，我们有什么数据呢？如图 3-1 所示。


图 3-1

以后我们还会有什么数据呢？

不知道！

在科技如此发达的今天，大概你仰望一下天空，卫星就会记录你的“仰天数据”吧！这个数据有什么用呢？就好像大数据有什么用呢？我们不知道，但是我们确信：一定会有用的！

## 3.2 大数据与传统数据

在这个人人都说大数据的时代，许多人对大数据的印象只是停留在仰望的阶段，其实大数据没人们说得那么神奇、玄乎或者是无所不能，今天我们就以传统数据作为比对，看看大数据究竟有什么特点让其处于时代的浪潮之巅。

大数据与传统数据相比的主要特点可以概括为：数据量“大”、数据类型“复杂”、数据价值“无限”（图 3-2）。

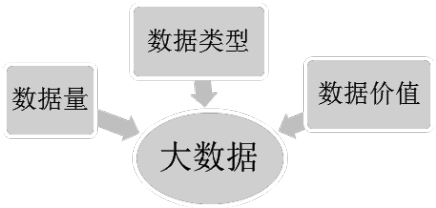



图 3-2

数据量大十分好理解，以前我们存储数据使用的单位是 KB，一个 Excel 表格也就几十到几百 KB，现在我们经常说到 GB 甚至是 TB 乃至 PB 的数据量级，它们的数量关系如下所示。

1MB=1024KB

1GB=1024MB

1TB=1024GB

1PB=1024TB

更直观一点，1KB 相当于 512 个汉字，1MB 就相当于六本红楼梦的字数……而淘宝网在 2015 年 3 月每天大约能产生 7TB 的数据量，相当于 4000 万本红楼梦的数据量，而中国最大的图书馆中国国家图书馆的藏书量是 3000 万册。由此看来，我们的大数据着实是数据量巨大了。而只说能够产生如此大量数据的原因有哪些呢？我们不妨从数据获取的方式、数据传输的方式和数据存储的方式来探讨数据量大的这个问题。

数据获取方式的质变是大数据能够产生的核心要素。传统的数据获取方式多是以人工的方式获取数据，最大的特点是手动输入数据，曾有一段时间，超市是通过要求收银员键入用户特征来采集用户数据的，键盘的样子大体上会是如图 3-3 所示的造型。

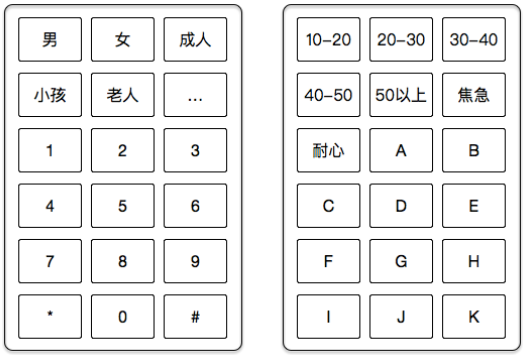



图 3-3

超市通过这样的方式来收集用户的数据，对收集的数据进行分析，来对用户画像与人群定位。试想在超市每天如此大的接待量情况下，收银员能否保证数据录入的准确性呢？与此同时，通过人工输入的方式每天能够采集多少数据呢？类似的这种键盘记录的方式还有许多人工录入数据的方式不再一一举例，传统记录数据的方式必定只能是小范围的，少量的和准确度欠佳的。而现在的数据获取方式大多是通过 URL 传输和 API 接口，大体上数据获取的方式有这样几类：爬虫抓取、用户留存、用户上传、数据交易和数据共享（图 3-4）。

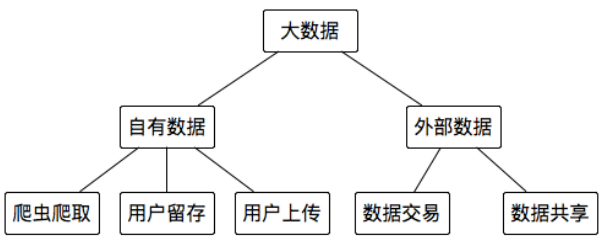



图 3-4

自有数据与外部数据是数据获取的两个主要渠道。在自有数据中，我们可以通过一些爬虫软件有目的的定向爬取，比如爬取一批用户的微博关注数据，某汽车论坛的各型号汽车的报价等。用户留

存多是用户使用了公司的产品或是业务，用户在使用产品或是业务中会留下一系列行为数据，这个构成了我们的数据库主体，通常的数据分析多基于用户留存的数据。用户上传数据诸如持证自拍照、通讯录、历史通话详单等需要用户主动授权提供的数据，这类数据往往是业务运作中的关键数据。相较于自有数据获取，外部数据的获取方式简单许多，绝大多数都是基于 API 接口的传输，也有少量的数据采用线下交易以表格或文件的形式线下传输。此类数据要么采用明码标价一条数据多少钱，或是进行数据共享，交易双方承诺数据共享，谋求共同发展。

至此，我们看到新时代的数据获取形式相较于传统数据获取的方式更加多元、更加高效。

同样的大数据与传统数据的传输方式也截然不同。传统数据要么以线下传统文件的方式，要么以邮件或是第三方软件进行传输，而随着 API 接口的成熟和普及就好像以前的手机充电接口，从千奇百怪、五花八门到今天的两大主要类别：iPhone 系统与 Android 系统。API 接口也随着时代的发展逐渐标准化、统一化，一个程序员只用两天的时间就能完成一个 API 接口开发，而 API 接口传输数据的效率更是能够达到毫秒级。

在数据存储方面，大数据的存储环境相较于传统数据的存储已经跃升了好几个数量级。犹记得十多年前软盘还非常高级，存储量达到 20MB 的软盘已然很贵，更别说 U 盘和移动硬盘了。

大数据与传统数据的另一个显著差异是数据类型的丰富。传统数据更注重于对象的描述，而大数据更倾向与对数据过程的记录。为了便于大家理解，下面简单的举个例子说明传统数据与大数据的记录方式有何区别。

传统数据的记录方式如表 3-1 所示。

表 3-1

| 姓　名 | 类　型 | 食　物 |
| --- | --- | --- |
| 小明 | 早餐 | 阳春面 |
| 小明 | 中餐 | 牛肉面 |
| 小明 | 晚餐 | 刀削面 |

大数据的记录方式如表 3-2 所示。

表 3-2

| 姓名 | 入场时间 | 入场方式 | 坐的位置 | 点餐时间 | 吃饭时间 | 吃饭内容 | 离开时间 |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 小明 | 9：23：34 | 自行车 | A2 | 9：28：42 | 9：30：18 | 阳春面 | 9：45：10 |
| 小明 | 12：25：21 | 步行 | A3 | 12：30：36 | 12：45：51 | 牛肉面 | 12：57：02 |
| 小明 | 19：02：12 | 自行车 | A2 | 19：15：27 | 19：25：04 | 刀削面 | 19：46：44 |

很明显地看到，传统数据和大数据记录数据的最大区别是大数据不仅对对象进行了描述，还加入了时间、地点等维度，这样的数据记录的是一个过程，从小明进入餐厅之前开始一直到小明离开餐厅，这整个过程都会被记录下来。而传统数据的记录方式更倾向于对结果的简单描述。

当然，大数据能记录的用户就餐数据远不局限于上述所列的字段，理想状况的大数据监控甚至会记录用户吃饭的方式、吃饭时的行为、吃饭时的面部表情等一系列数据，这些数据反映了用户对就餐环境的感受，对餐食口味的反应，进一步可以用来改进就餐环境、食物口味，给出点餐建议。

大数据与传统数据的核心差异在于其价值的不可估量。传统数据的价值体现在信息传递与表征，是对现象的描述与反馈，让人通过数据去了解数据。而大数据是对现象发生过程的全记录，通过数据不仅能够了解对象，还能分析对象，掌握对象运作的规律，挖掘

对象内部的结构与特点，甚至能了解对象自己都不知道的信息。

诸如某百科对一个人的描述与概括，记录了这个人的身高、体重、出生年月、兴趣爱好、日常活动、亲朋好友等数据，这些算是传统数据，通过这些传统数据你能知道和认识这个人。如果用大数据的方式来记录一个人，那就可以详细到他几点起床、睡眠质量、身体状况、每个时间点在做什么事等一系列过程数据，通过这些过程数据我们不仅知道和认识这个人，还能知道他的习惯性格，甚至能挖掘出隐藏在生活习惯中的情绪与内心活动等信息。这些都是传统数据所无法体现的，也是大数据承载信息的丰富之处，在丰富的信息背后隐藏着巨大的价值，这些价值甚至能帮助人们达到“所思即所得”的境界。

大数据价值的特殊之处就在于它的可挖掘性，同样的一堆数据，不同的人能得到不同层次的东西。就好像同样见一个人，有些人只看他的外貌好不好看，有些人能从他的表情中读出心理活动，从眼神中看出阅历，从衣着打扮中读出品味，从鞋子上读出生活习惯。而这些深层次的非表象的内容需要技巧与实力去挖掘出来，这就是我们说的数据分析与数据挖掘。

## 3.3 大数据在说什么

众所周知的啤酒与尿布就是一个典型的大数据应用实例，主要就是通过对大量数据做相关性分析得出买啤酒的人通常也会买尿布，于是就把啤酒和尿布放在了一起，在方便用户购物的同时也提高了销量。其实这种相关性在日常生活中随处可见，而且我们经常会在无意之间使用它们。

有一段时间，中午吃饭的时候我会和朋友玩几局狼人游戏，我们发现了一个有趣的现象，有的人在拿到狼人牌的时候总是做一些习惯性的动作，比如瞬间扫视全场或是看到牌之后摸一下鼻子，而

在他拿到好身份的牌时就不会出现这样的表现。这其实可以理解为一个相关性分析，玩家的动作与表情是一组数据，他拿到的牌是另外一组数据，通过比较它们的相关性能在一定程度上猜测这个人拿到的身份牌是什么。假设这样一个理想情况，玩家玩了无数局的狼人游戏，我们记录了他拿到不同身份牌时的表情、动作和眼神，下面截取一些数据作为示例如表 3-3 所示。

表 3-3

| 序　号 | 表　情 | 动　作 | 眼　神 | 表　现 | 身　份 |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 3 | 1 |
| 2 | 1 | 2 | 2 | 5 | 1 |
| 3 | 2 | 1 | 1 | 4 | 1 |
| 4 | 2 | 2 | 2 | 6 | 2 |
| 5 | 3 | 1 | 1 | 5 | 1 |
| 6 | 3 | 2 | 2 | 7 | 2 |
| 7 | 4 | 1 | 1 | 6 | 2 |
| 8 | 4 | 2 | 2 | 8 | 2 |

为了便于分析，我们把玩家不同的表情、动作、眼神用数字代替，身份为 1 表示好身份，身份为 2 代表坏身份。

我们看玩家的表情和身份的相关性，是不是能看出来表情数值越大用户越有可能是坏人？但是还是不准确，在玩家表情类型为 2 和 3 时的游戏中，拿过好身份牌也拿过坏身份的牌，但是我们相信这样一个趋势：表情数值越大越有可能是一个坏身份的人。这就是表情和身份数据的相关性，这个相关性能在一定程度上帮助我们判断这个人的身份是好的还是坏的。

那么进一步看，我们把用户的表情数值、动作数值与眼神数值求和得到“表现数值”，看一看“表现数值”和“身份”的关系，能否看出规律呢？我想你一定发现了：玩家表现数值小于等于 5 时是

坏身份，玩家身份大于等于 6 时是坏身份。得到这个结论后我们就能通过玩家的形体数据来判断这个玩家的身份了，是不是很神奇？

当然，这是一种理想情况，真实环境中数据不可能这么整齐规律，结果也不可能这么明显。这里只是想让大家知道大数据可以通过记录一系列数据（身体数据）来得到我们不知道的结果（你的身份牌）。

假设这样一种情况：有一款在线狼人游戏，大家是通过视频的形式在线游戏，每一局选定完角色后摄像头都会实时记录你的表情和行为数据，在经历了 $N$ 局游戏之后我们得到了大量的数据。由于互联网信息记录的便捷性，我们把 100 万在线玩家的数据汇集到了一起，保存玩家的视频数据，亲爱的同学们，我们能用这些数据做什么呢？这里不妨留作一个思考题给大家，心情发挥你们的想象力吧！

# 第4章 数据分析与数据挖掘

真正决定我们的，不是我们的能力，而是我们的选择！

## 4.1 分析与挖掘

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，认为数据挖掘就是数据分析，同时还与机器学习混淆在一起。严格意义上来说，数据分析和数据挖掘，作为两种不同的工作内容，数据分析更加偏向于业务，而数据挖掘更偏向于算法。通俗来说，数据分析是基于公司日常业务的观察、监测、分析与优化，而数据挖掘是基于数据库已有数据使用各种数据挖掘算法进行深度挖掘与讨论，同时机器学习算是数据挖掘的一个分支，隶属于数据挖掘的一部分。

如果我们直观地从分析和挖掘的字面上来理解：分析更像是对已有对象的全面描述、刻画、梳理后得出结论，而挖掘更倾向于对对象的解剖、分解、透视，发现不为人知的价值。就好像那里有一摊沙子，分析更倾向于分析这一摊沙子的颜色、结构、用途，而挖掘是给你一把铲子，把沙子挖起来，发掘沙子里面的东西。

如果用一句话概括数据分析的定义，那就是：借助数据来指导决策，而不是拍脑袋！其实传统行业的决策过多依赖于领导人的眼光和洞察力，而数据分析要做的事，就是把这些眼光和洞察力转化为人人可读的数字。我们可以粗略地把数据分析分为这样几个模块：明确分析目标、数据收集、数据清理、数据分析、数据报告、执行与反馈。下面针对这些不同的模块我们逐一阐述。

首先是数据分析的目的性极强，区别于数据挖掘的找关联、做分类、搞聚类，数据分析更倾向于解决现实中业务上的问题。我想解决什么问题？通过这次的分析能让我产生什么决策？比如是否在某个高校举办一场活动，是否把我们的补贴政策再增加 10 元，能否达到一个更好的效果？这样的决策和优化全部都是依赖于真实数据

之上的。达到目的是数据分析的核心目标。

其次是数据收集的方法，数据分析区别于数据挖掘的很重要的一点就是数据来源。数据分析的数据可能来源于各种渠道，数据库、信息采集表、走访等各种形式的数据，只要是和分析目标相关，都可以收集。而数据挖掘则偏向于数据库数据的读取。

与此同时，由于数据分析的数据来源较数据挖掘是直接从数据库调取更加杂乱无章，你可能是从别人的分析报告里找数据，从百度上搜索数据，这些数据的格式、字段都不统一，在这里你需要根据你的目的进行归类、整合、预估与填补等。

数据分析是全流程最重要的过程。这里最重要的事情是：时刻想着你的目标是什么？比如了解某个时间段的交易状况，你要根据这个目标做同比、环比分析等。这一块的方法极多，内容极大。而数据挖掘更倾向于使用贝叶斯、决策树、聚类分类等几个算法进行数据操作。

最后还要强调数据报告对数据分析的重要性。数据报告就是阐述你的结果，你可以搞一堆大家看不懂的公式证明你的专业性，但是这里需要你用最通俗易懂的语言告诉你的领导：做这件事有××%的概率收获××元。就是这么简单。

总体来说，数据分析更像一把枪，指哪打哪儿，非常实用，而数据挖掘更像武器研究，高科技自动化，前期投入高，时间跨度长，产出也很可观，数据领域的高精尖产品。前者偏向于业务，后者偏向于技术。前者在业务层次变现很快而且门槛较低；后者工资比较高，但是门槛也很高（图 4-1）。

数据分析与数据挖掘算是从数据出发的两个不同的分支，每个分支都有自己的特色。同时它们也有一些相似之处，诸如要求对数据库逻辑的了解、对数据结构的把控与对逻辑思维的要求等都有许多共通之处。

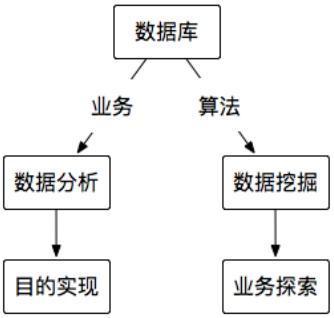



图 4-1

在公司运转的过程中，对数据分析与数据挖掘的需求是持续不断的，两者同时基于数据数据库对公司的各项业务给予建议和优化方案，相辅相成，在一家成熟的公司中这二者都是不可或缺，同样重要。

## 4.2 选择自己的路

在我们了解数据挖掘和数据分析之后就会面临一个非常实际的问题：我该选择数据分析还是数据挖掘？或者是先做数据分析然后再做数据挖掘或是先做数据挖掘再做数据分析或是两者都做？我们不妨结合目前的市场情况对数据分析和数据挖掘有一个直观的认识。

数据分析和数据挖掘算是随着互联网行业的蓬勃发展而产生的相对新兴的行业，绝大多数的数据分析或是数据挖掘从业者都是半路出家或是自学成才。许多人之前都是公司业务员出身，直到发现日常的工作需要大量的数据指导时才萌发了做数据分析的想法。可以说数据行业是从业务转变而来，当我们需要它时，它就出现了。

对于绝大多数人来说，在入门这一行业时都是被大数据时代这个概念吸引而来，我们从对大数据的盲目崇拜，到基本了解数据分析和数据挖掘，之后又会听到许多人谈论机器学习，了解到数据建模。而真正开始数据分析工作之后，才发现日常工作的绝大多数都

是由 Excel 完成，每天上班开始进行 Excel 表格制作、需求处理、数据透视表然后掌握一些诸如 Vlookup 函数就能基本满足日常工作的需求。在度过了一开始的兴奋期之后任何人都不免有些灰心丧气，本来以为是高大上的大数据分析行业，现在却沦落成了“表哥”、“表姐”，想象中的数据分析工作不是这样的，大家开始心有不甘去尝试一些“高大上”的功能和技法。

大家在工作之余会开始去了解数据分析与数据挖掘需要掌握的核心技能是什么，毕竟 Excel 不能算作数据分析的核心技能。就像所有行业刚刚兴起的那样，网络教程和指导鱼龙混杂，一个月学会数据分析或是三个月精通数据分析的教程遍地都是，见猎心喜的状态出现在每个人的脸上。接下来大家就会开始面临“R or Python”这样一个问题？SAS 自然是不愿意学习的，SAS 软件的安装方法就能阻止一大半的自学者。SPSS 是不愿意学习的，点点鼠标的事情看起来还是没有技术含量……

下面结合笔者的自身经历来谈一下，曾经有那么一段时间在数据分析工作上也遇到了瓶颈，也曾问过自己数据分析行业的核心竞争力是什么？能不能找一些专业靠谱的技巧写在简历上，诸如精通 R 语言，熟练掌握 Python 等。先后购买过许多关于 R 语言与 Python 相关的书籍教程，也在某问答类网站上学习探索。经过一番挣扎之后收效甚微，到头来才发现没有应用场景的自我学习只能停留在皮毛阶段，自学数据挖掘的关键是构建应用场景。

在近一个月的探索与学习后对 R 语言与 Python 有一个大致的了解和认识，自己尝试用 Python 写了一个爬虫，虽然问题颇多，但也算是在数据挖掘上行进了一小步。之后又做了基于 R 语言的 Logistic 回归建模，模型初见成效也增加了自己对数据挖掘的信心。然而随着时间的推移，逐渐接触决策树、神经网络、贝叶斯算法才发现自己统计学知识的缺乏。在接下来的一段时间从概率与分布开始了统计学的二次学习，对最小二乘估计等方法从最基本的逻辑开始梳理，在之后是回归、分类、聚类、贝叶斯……一步一个脚印把统计学的

这些硬骨头“啃”下来，算是打牢了数据挖掘理论上的基础。这里不妨为尝试学习数据挖掘理论的同学梳理一下数据挖掘各个层级的基础知识，数据挖掘的理论知识体系大抵如图 4-2 所示。

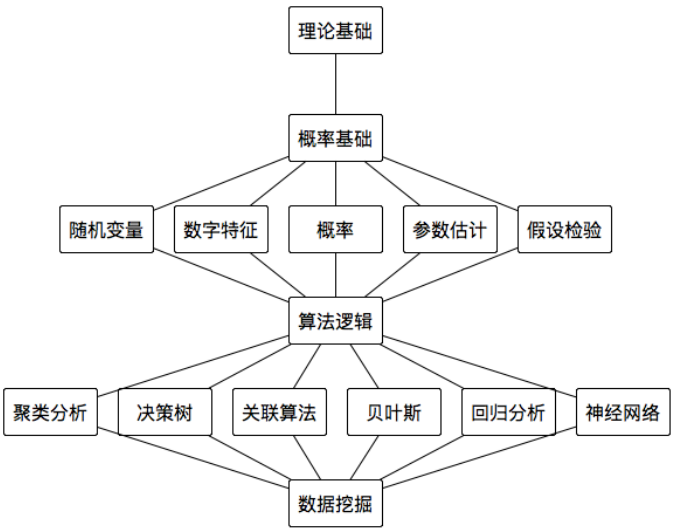



图 4-2

每一个分支都对应着繁杂的数学推延和逻辑，需要坚实的数学基础作为支撑才能持续学习下去。按道理说到这里自己应该再次推进数据挖掘的工作了，但是考虑到之前的问题：理论学习需要应用场景来巩固学习。与此同时，数据挖掘的另外一个要求是对编程的熟练掌握，这又将是一个漫长的过程，而且数据挖掘的应用场景远小于数据分析场景，因此，笔者目前依旧处于数据分析岗位。

度过了那样一个阶段后，笔者现在潜心于研究数据分析，用数据的方法解决公司实际运营中遇到的种种问题，变现速度很快。虽然笔者目前在日常工作中遇到的绝大多数问题还是在 Excel 中进行处理，但是已经没有当初那种急躁和疲惫的情绪，绕了一大圈又回归到最初的选择。

从数据分析入门开始，许多同学在数据分析岗位工作一段时间

之后都会遇到瓶颈期，这个时候会迷茫犹豫，如何跳脱现在的状态？这时候我们会发现数据挖掘的高大上、高精尖和高薪资，有人因此走上数据挖掘的道路。也有人发现数据挖掘对算法与编程的要求极高，作为非科班出身想要自学实在是困难无比，转头继续做数据分析。每个人都有自己不同的选择，每个人都会经历这样一个阶段，无论做出怎样的选择都希望大家能坚信自己的选择，坚持自己的看法，在数据分析或是数据挖掘的道路上取得自己的成功。

# 第5章 如何做好数据分析

“数据分析入门简单吗？”

“简单！”

“人人都能做数据分析师吗？”

“未必！”

## 5.1 数据分析

想要学习数据分析的你一定从各种渠道了解过如何学习数据分析，从某些问答类网站得到的答案大多是：

统计学：《统计学》《商务与经济统计》

软件：SAS、SPSS、R、Python、MySQL

书籍：《R 语言实战》《Python 基础教程》《Python 核心教程》《概率论》

……

你会开始买书、下载软件、开始学习“Hello World！”、逛论坛……

新世界的大门就此打开！

但是，

能坚持一周的人有几个呢？

能坚持一个月的人又有几个呢？

能坚持三个月的人又有几个呢？

相信我，坚持到底的那个人不会是你！

100 个人如果有 1 个人坚持到底，我并不会称赞他，只能说那个人固执得很！

为什么？

人都是有惰性的，你有、我有、大家有！我们必须首先承认自己的惰性再来谈如何克服惰性。

从心理学角度来说，克服惰性的意志力来源于努力带来的回馈，就好像你愿意用尽全力去砸开一个核桃是因为你可以吃到果肉。如果我不给你核桃而是给你两本书，让你学习如何砸核桃，你能看多久呢？同样的脱离实际问题的学习数据分析就好像让一个从来没见过核桃的人学习如何砸核桃，你很努力地在强迫自己学习却不知道能取得怎样的效果，一切都是依托于别人描绘的数据分析的未来是多么广阔，前途是多么光明。而这些画面的本身就有些画饼的成分，短时间内肾上腺激素的分泌刺激你去狂热地学习，但是这种狂热能够持续多久呢？只有认识到自己的惰性才能克服自己的惰性，设立必要的回馈机制才能让自己在一件事情上持续地坚持下来。

既然说了这么多，那到底该怎么学习数据分析，如何克服自己的惰性呢？我们先来了解一下数据分析的主要工作内容。

（1）制作报表

（2）异常数据分析

（3）数据需求

（4）项目性分析

接下来我们将通过四个小节（5.2 节、5.3 节、5.5 节和 5.6 节）逐步为大家介绍这四个模块，让我们一起开始数据分析的探索之旅。

## 5.2 制作报表

我们先来聊聊制作报表，先不谈仪表盘、水晶报表、可视化方法，咱们就来聊聊 Excel。任何数据分析工作都是从制作 Excel 报表开始的，我相信绝大多数求职者都会在自己的简历上写熟练运用 Word、Excel、PPT 办公软件，同时绝大多数人都不知道 Word、Excel、

PPT 的工具有多么强大，功能有多么复杂。我使用 Excel 做数据分析这么多年，也只能说能够熟练运用 Excel，对于 Word 和 PPT 的掌握真的只能局限在一知半解，仅仅只算做了解它们。我们先谈谈在数据分析时常用的几个 Excel 功能。

首先是数据透视表和数据透视图，作为一个数据分析人员这个要求简直是基础得不能再基础了，如果不了解数据透视表的人来做数据分析就好像分不清基本颜色就去做设计，这个是最低要求，相信绝大多数的小伙伴都会，也就不做赘述了。

接下来说一说我们常说的描述性统计变量。描述性统计变量是我们刻画和描述一个数据样本的最基本手段，下面的这几个函数大家会经常用到。如表 5-1 所示。

表 5-1

| 函数类别 | 函数名称 |
| --- | --- |
| 求和函数 | Sum()　Sumif()　Sumifs() |
| 计数函数 | Count()　Countif()　Countifs() |
| 平均值 | Average() |
| 中位数 | Median() |
| 最大值 | Max() |
| 最小值 | Min() |
| 方差 | Var() |

如果把一组数据比作一个三维物体，求和与计数用来衡量它的长、宽、高，平均数用来衡量它的密度，中位数用来衡量它的几何中心，最大值与最小值用来衡量它的突出、凹陷，方差用来衡量它是否均匀……所以我们把上面的几个基本的函数叫作描述性统计变量，我们通过这些变量将一堆抽象的数据描述得直观和便于理解。

接下来我们聊一聊统计制图，作图这件事情说简单也简单，往深了讲也能说出三大要素五大标准，但是制作图表的关键一定要明白：

**图表是一种表达方式**

制作图表的核心标准是：

**受众能迅速准确地获得你想表达的内容**

当然，市场上流传着麦肯锡咨询公司的图表制作方法，四大事务所的图表制作准则等，它们的确好看，但是我们必须明白那是人家在信息传递的基础上做到极致，不仅把一件事情表述得十分清晰，同时能在细节上传递出严谨和极致的态度。我们在学习这些高大上的标准和方法之前一定要先学习如何把结果清晰地呈现，让别人更加直观地感受到你想要传递的内容。这里有一个简单的标准供参考：

- 折线图传递变化趋势的信息
- 饼状图传递组成成分的信息
- 柱状图传递数值大小的信息
- 散点图传递数据集中度的信息
- 面积图传递数值累积的信息

上面的这五种图表基本上可以满足日常需求，顺便运用上述规则进行交叉混合可以让你在图表制作上更上一个台阶。下面举一些样例：

（1）我想了解公司 2015 年每个月用户的增长情况，如图 5-1 所示，用哪种图表更合适？

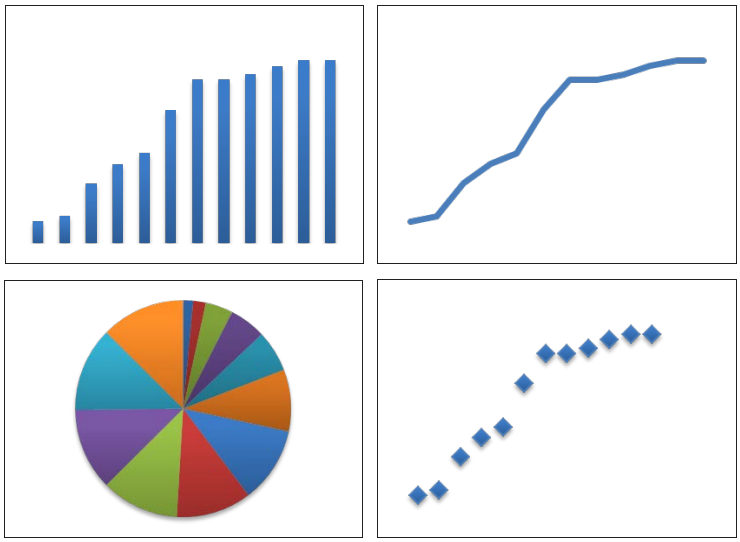


图 5-1

（2）我想了解公司 2015 年 A 业务与 B 业务的交易额，如图 5-2 所示，用哪种图表更合适？

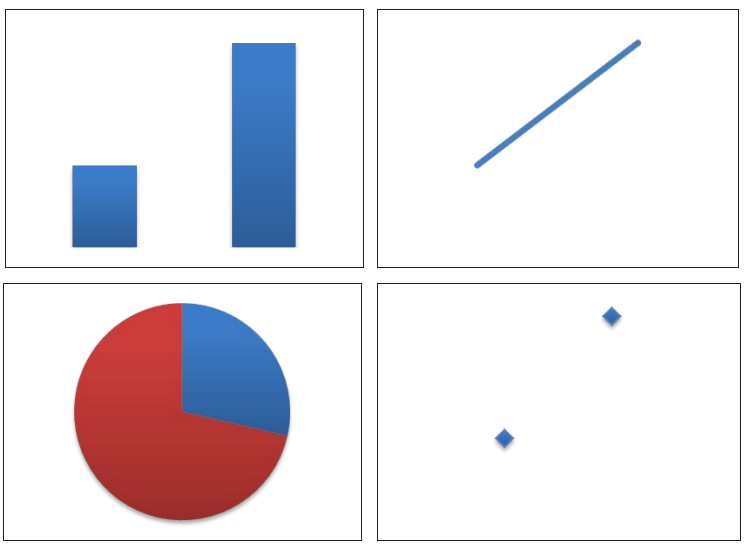


图 5-2

（3）我想了解 2015 年产品线各个业务的数据表现，如图 5-3 所

示，用哪种图表更合适？

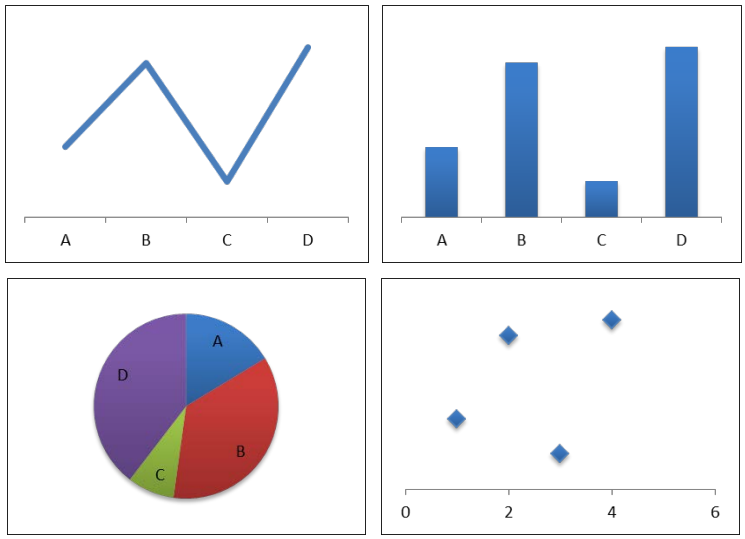



图 5-3

（4）我想了解公司 2015 年 Top 20 用户的交易额，如图 5-4 所示，用哪种图表看更合适？

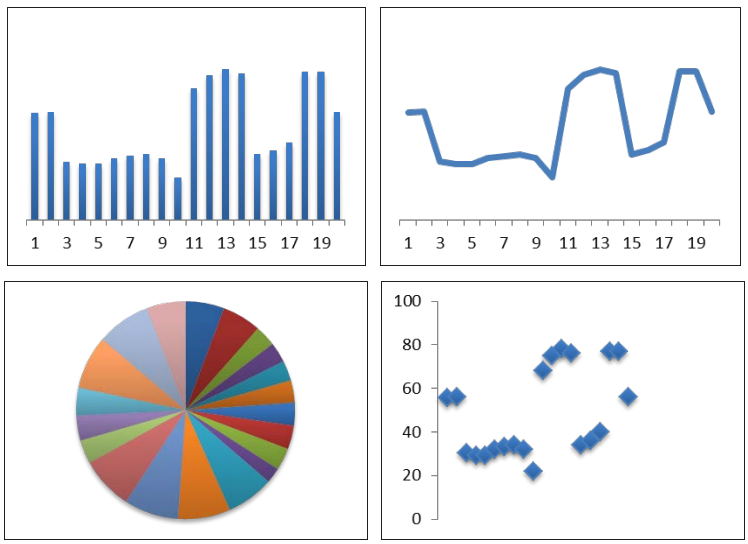



图 5-4

上面列举了四个简单的问题，每个问题都可以用 Excel 的基本图表来解决问题，每个问题都是数据分析日常中经常面临的问题，针对每一个问题我都只做了柱状图、折线图、饼状图、散点图，大家可以分门别类地看下针对每一个问题哪一张图表最能解决问题，最能让提问者清晰直观地感受到数据表现。

- 第一个问题倾向于对 2015 年每个月用户增长的趋势进行呈现，所以折线图是最优解。
- 第二个问题倾向于对 A 与 B 两个业务的数值大小比较，所以柱状图是最优解。
- 第三个问题倾向于整体业务中各个业务的占比，所以饼状图是最优解。
- 第四个问题倾向于 Top20 用户的整体分布与聚集情况，所以散点图是最优解。

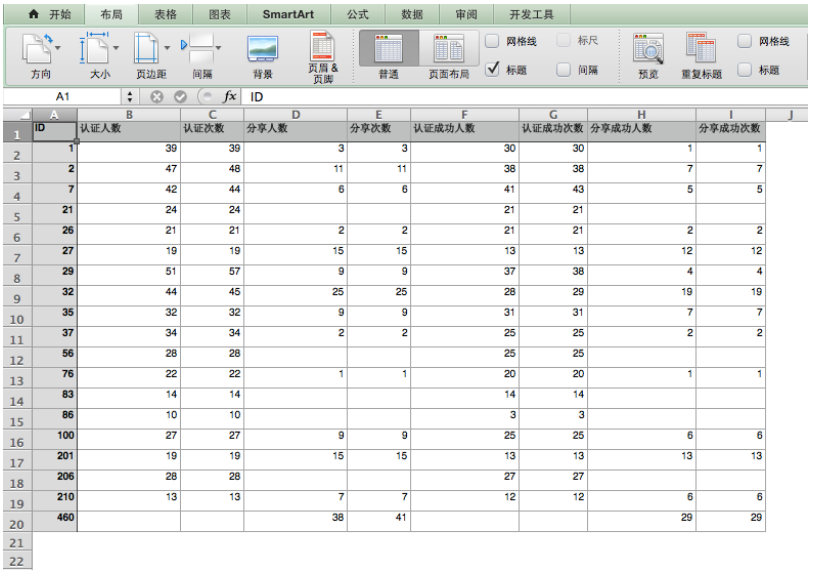
通过简单的四个案例，不知道是否让大家对柱状图、折线图、饼状图和散点图有进一步的认识，实际应用中柱状图往往与折线图制作复合图表，饼状图也会伴随着复合饼图，散点图有时候会以气泡图的形式呈现更多信息。复杂的图表大多是以这四种图表为基础进行拼接扩充，这些简单的原则在熟练运用之后会使你的图表制作水平显著提升。Excel 图表算是大家接触可视化的第一步，随着简单图表的制作逐渐无法满足日常需求，或者是想在可视化的道路上精益求精，会接触到仪表盘、图表控件、可视化工具等。永远记住图表制作的核心标准：

**受众能迅速准确地获得你想表达的内容**

知道每一个图表的特点、优势、特征，使用它们清晰地表述你的观点。

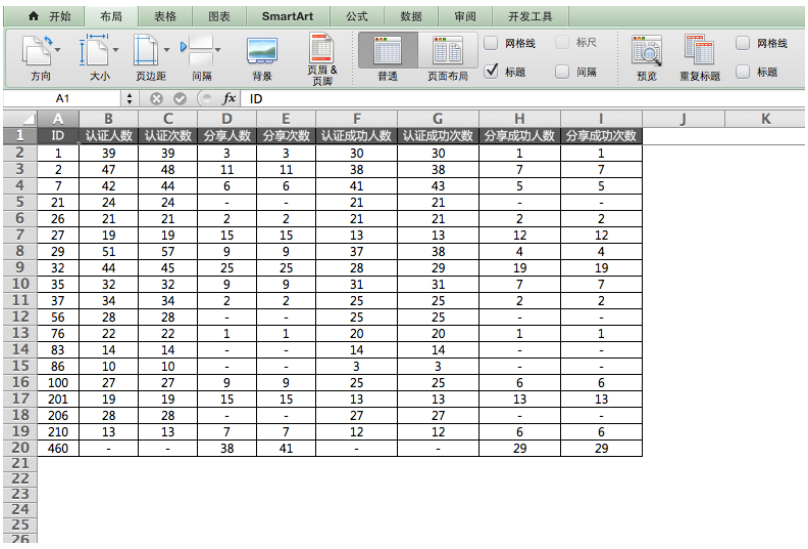
作图是制作报表的一个重要环节，与此同时，近乎“变态”的严谨性是制作报表的最基本要求之一，比如报表制作中的微软雅黑 9 号字体，比如字段行底色浅色 35%灰色底色，字体白色加粗，上

下居中，左右居中，首行首列做冻结窗格等，都是一张报表专业性的体现，比如下面两张图表（图 5-5 和图 5-6）。



| ID | 认证人数 | 认证次数 | 分享人数 | 分享次数 | 认证成功人数 | 认证成功次数 | 分享成功人数 | 分享成功次数 |
|---|---|---|---|---|---|---|---|---|
| 1 | 39 | 39 | 3 | 3 | 30 | 30 | 1 | 1 |
| 2 | 47 | 48 | 11 | 11 | 38 | 38 | 7 | 7 |
| 7 | 42 | 44 | 6 | 6 | 41 | 43 | 5 | 5 |
| 21 | 24 | 24 | | | 21 | 21 | | |
| 26 | 21 | 21 | 2 | 2 | 21 | 21 | 2 | 2 |
| 27 | 19 | 19 | 15 | 15 | 13 | 13 | 12 | 12 |
| 29 | 51 | 57 | 9 | 9 | 37 | 38 | 4 | 4 |
| 32 | 44 | 45 | 25 | 25 | 28 | 29 | 19 | 19 |
| 35 | 32 | 32 | 9 | 9 | 31 | 31 | 7 | 7 |
| 37 | 34 | 34 | 2 | 2 | 25 | 25 | 2 | 2 |
| 56 | 28 | 28 | | | 25 | 25 | | |
| 76 | 22 | 22 | 1 | 1 | 20 | 20 | 1 | 1 |
| 83 | 14 | 14 | | | 14 | 14 | | |
| 86 | 10 | 10 | | | 3 | 3 | | |
| 100 | 27 | 27 | 9 | 9 | 25 | 25 | 6 | 6 |
| 201 | 19 | 19 | 15 | 15 | 13 | 13 | 13 | 13 |
| 206 | 28 | 28 | | | 27 | 27 | | |
| 210 | 13 | 13 | 7 | 7 | 12 | 12 | 6 | 6 |
| 460 | | | 38 | 41 | | | 29 | 29 |

图 5-5



| ID | 认证人数 | 认证次数 | 分享人数 | 分享次数 | 认证成功人数 | 认证成功次数 | 分享成功人数 | 分享成功次数 |
|---|---|---|---|---|---|---|---|---|
| 1 | 39 | 39 | 3 | 3 | 30 | 30 | 1 | 1 |
| 2 | 47 | 48 | 11 | 11 | 38 | 38 | 7 | 7 |
| 7 | 42 | 44 | 6 | 6 | 41 | 43 | 5 | 5 |
| 21 | 24 | 24 | - | - | 21 | 21 | - | - |
| 26 | 21 | 21 | 2 | 2 | 21 | 21 | 2 | 2 |
| 27 | 19 | 19 | 15 | 15 | 13 | 13 | 12 | 12 |
| 29 | 51 | 57 | 9 | 9 | 37 | 38 | 4 | 4 |
| 32 | 44 | 45 | 25 | 25 | 28 | 29 | 19 | 19 |
| 35 | 32 | 32 | 9 | 9 | 31 | 31 | 7 | 7 |
| 37 | 34 | 34 | 2 | 2 | 25 | 25 | 2 | 2 |
| 56 | 28 | 28 | - | - | 25 | 25 | - | - |
| 76 | 22 | 22 | 1 | 1 | 20 | 20 | 1 | 1 |
| 83 | 14 | 14 | - | - | 14 | 14 | - | - |
| 86 | 10 | 10 | - | - | 3 | 3 | - | - |
| 100 | 27 | 27 | 9 | 9 | 25 | 25 | 6 | 6 |
| 201 | 19 | 19 | 15 | 15 | 13 | 13 | 13 | 13 |
| 206 | 28 | 28 | - | - | 27 | 27 | - | - |
| 210 | 13 | 13 | 7 | 7 | 12 | 12 | 6 | 6 |
| 460 | - | - | 38 | 41 | - | - | 29 | 29 |

图 5-6

图 5-5 是从数据库直接抓取并下载的数据，图 5-6 是经过简单加工处理的数据报表，大家可以简单对比一下这两张表格，随着字体调整、标题行突出、居中设置、缺失值处理等，图 5-6 的表格相较于图 5-5 的表格要专业许多。严谨和极致其实是数据分析从业人员的一种态度，就好像一个标点一个符号和一个大小写，这些良好的习惯如果不好好培养，总有一天会在数据上因为多一个 0 或是少一个 0 而犯错，这种错误往往是数据分析师的灾难，它可能直接影响到你的数据可信度。造成的后果小到被领导骂一顿扣绩效，大到丢掉这个饭碗。数据是万万不能出错的。其实我们经常被灌输的一个概念是 99%=0，对数据分析师的要求就是所有的数据必须是正确而准确的，一旦有一个错误那么这个报告就不能看了，这是原则性问题。数据大厦的每一个环节都是精确无误而又必不可少的，不像造房子，你可以在薄弱的地方多加固一些，不像制订计划，A 计划完成不行还有 B 计划，数据错了就是错了，没有挽回的余地，没有让步的余地。在数据的领域里只有完全正确与错误，不存在模棱两可，不存在“probably”（可能）。

所以说极致和严谨是一种习惯和态度，应该居中对齐的地方就要全部居中对齐，需要左对齐的地方就要左对齐。所有图表的长为 12 厘米，高为 6 厘米必须统一。但是与此同时不得不说有时候这些数据格式的调整和统一规范需要耗费许多时间，这时候大家熟悉和了解宏和 VBA 就十分必要了！

Excel 开发工具中的宏的录制就不再赘述了，大多数读者都知道如何使用宏录制简单的操作，这里简单地说下 VBA 编程。

一切 VBA 编程的学习都是从宏开始的，宏是 VBA 编程的外在表现形式，宏让你能够无感知的参与 VBA 编程，但是要是想更好地理解和使用宏，就得要了解宏后面对应的逻辑，在逐渐熟悉了解 VBA 的过程中，用 VBA 优化和取代宏的工作，来帮助完善效率。抛开一切 VBA 的起源、原理，承上启下，不管它是面向对象还是结构化语言等专业术语，跟着我的步伐，让我们用十分钟的时间熟悉

和了解 VBA，十分钟 VBA 编程入门。

十分钟 VBA 时间

Sub & End Sub：常用的 VBA 语句都是以 Sub 宏名称()开始，以 End Sub 结束。随手在 Excel 里录制一个简单的宏，然后单击 Visual Basic 就能查看到你录制的宏对应的 VBA 代码，我们不管代码的内容是什么，开头一定是 Sub 宏 1()，结束为 End Sub，英文单引号“'”后面对应的是你的快捷键的注解（英文的单引号表示语句注释）。

Range. Select & Cells & Selection：就好像我们设置 Excel 里面某一行数字的字体之前要先选中这一行一样，在 VBA 里面同样需要选中一个范围，然后进行下一步操作。Range 为范围，Select 为执行的操作，Range.Select 就直接可以理解为选中 Excel 工作簿的一个范围。Range 只是一个概括性的说法，要使用 VBA 选中特定的范围我们需要使用程序能够读懂的语言 cells()，cells(1,1)为第一行第一列，cells(i,j)为第 i 行第 j 列，Range(cells(1,1), cells(i,j)).Select 就表示选择从(1,1)到(i,j)的矩形区域。

With & Font & Interior：在 Excel 中想要设置你选中目标的字体、字号、单元格属性的时候一般会进行右击选择字体或者段落，With 这个属性就相当于右击看一看你设置的属性，而 Font 就表示字体类设置，Interior 就表示单元格类的设置，所以我们会使用 With Selection.Font 来设置选择区域的字体，使用 With Selection.Interior 用来设置选择区域的单元格属性。

Bold&Color&Other：这些属性变量是我们在 VBA 使用过程中对选择区域对象的属性设置，诸如加粗、设置颜色、倾斜等，在以后的 VBA 使用过程中还会遇到其他属性设置，大家可以在使用的时候直接通过网络搜索，分分钟就能解决问题，在这里做一一罗列没有任何意义。

“=”&”><”&”if”：这些符号分别对应着赋值、判断和循环，对于这些操作主要还是逻辑思维的清晰度和编程能力关联不多，在

这几个方向上各类编程语言都是大同小异，这里也无需累赘。

简单说，VBA 的基本使用就是明确整体框架（Sub&End Sub），选择需要处理的范围（With & Font & Interior），进行相应的设置（Bold&Color&Other），使用逻辑贯穿始终（“=”&”><”&”if”），接下来举一个用 VBA 设置数据格式的简单的案例。

```
Sub 格式()
' 格式 宏
    Range(Cells(1, 1), Cells(20,20)).Select  ''选择单元格(1, 1)到(20,20)
    With Selection.Font        ''设置选中格式的字体
        .Name = "微软雅黑"      ''字体设置微软雅黑
        .Size = 9      ''字体设置为 9 号
    End With       ''结束设置
    Selection.Columns.AutoFit ''选择所有行自动适应行高
    Selection.Rows.AutoFit    ''选择所有列自动适应列宽
    Selection.Font.Color = RGB(255, 255, 255)
''字的颜色号白色
    Selection.Interior.Color = RGB(89, 89, 89)
''背景颜色号为 2
    Selection.Font.Bold = True        ''加粗
    Cells(1, 1).Select    ''选择单元格(1,1)
End Sub    ''结束
```

通过对 VBA 的简单了解让我们对宏有了进一步认识，通过对宏的熟练使用，能让我们在日常使用 Excel 的过程中更加高效而且准确地处理问题，很多重复性的工作都可以被宏或者是 VBA 所取代。

讨论完报表制作，大家可以感受到一个看似简单的数据报表的制作里面有很大的空间可以挖掘，数据报表的准确性是 100%还是 0，很多时候都是一念之间的事情，所以把报表制作做到极致，把极致的思想贯穿到日常工作的始终是一个合格的数据分析人员应该具有的态度和素质。

## 5.3 异常数据分析

报表制作的掌握与运用是数据分析最基本的要求，异常数据分析则是数据分析工作的第一步。异常数据分析只是一个概括性的说法，我们可以把所有不符合随机波动的数据概括为异常数据，异常数据分析也可以理解为找不同，然后分析不同。

一个公司的运作从市场推广开始，到用户覆盖，到用户入场、转化、活跃、交易、留存每一个环节都对应着各种数据统计和数据报表。异常数据分析的第一步就是要能够从这些众多的数据中发现异常数据，举一个简单的例子：

如图 5-7 所示为某公司 1 月 A~H 八个指标，大家不妨仔细观察这些数据中有什么异常或者是规律？

| 日期 | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1月1日 | 39 | 39 | 3 | 3 | 30 | 30 | 1 | 1 |
| 1月2日 | 47 | 48 | 11 | 11 | 38 | 38 | 7 | 7 |
| 1月3日 | 42 | 44 | 6 | 6 | 41 | 43 | 5 | 5 |
| 1月4日 | 24 | 24 | - | - | 21 | 21 | - | - |
| 1月5日 | 21 | 21 | 2 | 2 | 21 | 21 | 2 | 2 |
| 1月6日 | 19 | 19 | 15 | 15 | 13 | 13 | 12 | 12 |
| 1月7日 | 51 | 57 | 9 | 9 | 37 | 38 | 4 | 4 |
| 1月8日 | 44 | 45 | 25 | 25 | 28 | 29 | 19 | 19 |
| 1月9日 | 32 | 32 | 9 | 9 | 31 | 31 | 7 | 7 |
| 1月10日 | 34 | 34 | 2 | 2 | 25 | 25 | 2 | 2 |
| 1月11日 | 28 | 28 | - | - | 25 | 25 | - | - |
| 1月12日 | 22 | 22 | 1 | 1 | 20 | 20 | 1 | 1 |
| 1月13日 | 14 | 14 | - | - | 14 | 14 | - | - |
| 1月14日 | 10 | 10 | - | - | 3 | 3 | - | - |
| 1月15日 | 27 | 27 | 9 | 9 | 25 | 25 | 6 | 6 |
| 1月16日 | 19 | 19 | 15 | 15 | 13 | 13 | 13 | 13 |
| 1月17日 | 28 | 28 | - | - | 27 | 27 | - | - |
| 1月18日 | 13 | 13 | 7 | 7 | 12 | 12 | 6 | 6 |
| 1月19日 | - | - | 38 | 41 | - | - | 29 | 29 |
| 1月20日 | 39 | 39 | 3 | 3 | 30 | 30 | 1 | 1 |
| 1月21日 | 47 | 48 | 11 | 11 | 38 | 38 | 7 | 7 |
| 1月22日 | 42 | 44 | 6 | 6 | 41 | 43 | 5 | 5 |
| 1月23日 | 24 | 24 | - | - | 21 | 21 | - | - |
| 1月24日 | 21 | 21 | 2 | 2 | 21 | 21 | 2 | 2 |
| 1月25日 | 19 | 19 | 15 | 15 | 13 | 13 | 12 | 12 |
| 1月26日 | 51 | 57 | 9 | 9 | 37 | 38 | 4 | 4 |
| 1月27日 | 44 | 45 | 25 | 25 | 28 | 29 | 19 | 19 |
| 1月28日 | 32 | 32 | 9 | 9 | 31 | 31 | 7 | 7 |
| 1月29日 | 34 | 34 | 2 | 2 | 25 | 25 | 2 | 2 |
| 1月30日 | 28 | 28 | - | - | 25 | 25 | - | - |
| 1月31日 | 22 | 22 | 1 | 1 | 20 | 20 | 1 | 1 |

图 5-7

异常数据分析，顾名思义，先得发现异常，上面某公司 1 月份 31 天的八个指标详细地罗列了出来，一眼望去密密麻麻的全是数值，仔细看看有什么异常点或者规律呢？大家不妨开启“找茬”模式，先仔细观察，在心里罗列一下，尝试找出至少三个的异常点，看一看和下面的答案是否有重合和不同。

上面密密麻麻的数值看着一定很头晕，没有结合实际业务的数据观察着实枯燥无味，不知道大家有没有发现这样几个问题：

（1）A 与 B 两列除了 1 月 19 日缺失，其他日期全都有数值

（2）C 与 D，E 与 F，G 与 H，两两组合要么同时缺失，要么同时有数值

（3）1 月 20 日到 1 月 31 日的数值与 1 月 1 日到 1 月 12 日相同

（4）A 与 B，C 与 D，E 与 F，G 与 H 在 90%的情况下都是相同的

（5）1 月 14 日的各项指标明显低于其他日期

……

像上面的异常数据查找还可以有很多，每个人从不同的角度都可能有不同的发现，这里并没有标准答案。同样的，这里需要解释一下，异常数据并不是说这个数据不正常，由于业务的交叉性和相互影响，异常的性质有可能是规律的性质。就好像在啤酒和尿布的案例中每次啤酒的销量变好的时候尿布的销量也变好，当我们看到这两个产品的销量时这种规律性也可以被称为异常数据发现。

既然异常数据分析是从众多的数据中找出规律和不同，那这件事情是不是人人可为，把数据拿过来看就好了呢？回答这个问题的时候我们不妨回想一下刚刚的数据案例中，你从中发现了几个异常呢？这里就要提出一个数据敏感度的概念，什么是数据敏感度呢？数据敏感度可以类比成一个人的音乐感觉，有些人的乐感就是好，有些人的乐感就是差。数据敏感度就是一个人对数据的感觉，同样

的，有些人的数据敏感度高，有些人的数据敏感度低。

数据敏感度是一个人对数据的主观感觉，能帮助你从众多的数据中挑选出想要的信息，甄别出不一样的数据点，不妨做一个小测试，请在短时间内填补空缺的数值：

1，3，6，__ ，15，21

1，3，7，__ ，31，63

数字敏感度就是能在极短时间内看出空格中应该填什么数字的能力。这种能力能帮助你在众多的数据报表中快速发现异常，找到关键问题。

第一个空格的答案是 10，因为

3-1=2

6-3=3

21-15=6

你的第一直觉应该是 2、3、4、5、6，这就是数据敏感度。

第二个空格的答案是 13，因为

3-1=2

$7-3=4=2^2$

$63-31=32=2^5$

你的第一直觉应该是 2、$2^2$、$2^3$、$2^4$、$2^5$，这就是数据敏感度。

数据敏感度这种感觉在异常数据分析中必不可少，不可或缺。如果对每天众多的数据报表缺乏敏感度，缺乏异常识别和异常预警，最终只能沦为一个“做报表”的。

既然数据敏感度如此重要而又必不可少，那么数据敏感度能否培养呢？如何练就一双慧眼能在大段大段的数字中发现异常找到规

律呢？方法很简单，只有三个字：

**背数据**

相较于各类思维培训方式，数据阅读的方式，背数据绝对是简单粗暴而又十分高效的方式。就好像英语的语感是靠背文章来实现，数据的感觉也来源于背数据。

在我一开始从事数据分析工作的时候每晚都会被强制要求背记今天的交易数据、用户数据等，开始很不理解，后来慢慢上道了才豁然开朗。培养对数据的感觉太重要了。能让你一眼就看出哪里的数据有问题，哪里的数据有关联，然后开始下一步的分析。之前还遇到过导师让同学们背代码的事情，同样的几乎所有人对这件事情都表示排斥，代码和算法这些方法论与强逻辑关系的语言为什么要背记呢？老师是不是太迂腐了，有种回到科举制背记四书五经的感觉，其实从某种意义上来说，背代码不仅仅是背代码，导师是希望你能培养代码的感觉。

对于代码的感觉，对于数据的感觉，对于英语的感觉，对于音乐的感觉……谁能说出来那种感觉是什么？能写出来吗？不能！但是有用吗？你说呢？

让大家认可这件事情之前其实需要大家能够接受这样一件事情，并且重视这样一件事情。许多东西不是我们原先想象得那样，许多东西也不是我们以为的未来的那个模样，抛去思维定式，这是尝试接受新的东西的核心方法。

之前给高中生辅导数学，强制要求他们把定理抄下来。大家都会问这个问题：这样做有用吗？定理不是只要理解就好了的吗？一开始大家觉得没用。所有人都觉得没用，定理理解就好了！为什么要抄？但是许多时候只有抄下来才知道有些东西不是你想当然的理解了就懂了。学了多年的数学，教过好几个孩子，目前来看效果还算不错。

数据感觉的培养需要你强迫自己背记数据，用文科的方式去培养感觉，假设公司每天有十几张数据报表，从中间挑取几张关键的核心报表开始着手背数据，一天一天地坚持，给自己设立奖惩机制，争取做到记得住历史七天公司各个业务的每一个环节对应的数据，坚持一个星期，你会发现你对公司数据的理解度远超从前，你会发现接下来你不再需要背数据，每天的数据都在你的预期之内。当你能熟练记忆公司各项业务的各个环节的历史七天数据，你对公司明天的数据表现就会产生一个心理预期值，这个预期值就来源于你对数据的感觉，它的准确度在短期预测中会远超时间序列或者是回归分析，人脑的复杂度在这方面的表现会令你感到惊讶。同样，若英语的语感培养得好，英语的填空题、选择题的答案许多时候不需要去思考语法规则、是否呼应上下文、过去式将来时等，答案就在你看到那个空格的瞬间出现在脑子里，这就是语感。当有一天你能十分自信地凭着感觉预测公司未来几天各项数据在各个环节的表现时，你的数据感觉算是培养得初见成效了。这个时候你再去看数据报表，一旦某些数值和你的预期不一样，那么这个数据就极有可能是有问题了。换一个角度来说，数据敏感度也是由熟能生巧带来的。你对数据的熟练度足够高，所以能够帮助你一眼看穿数据的本质，直击数据背后隐藏的问题。

我们不妨做一个简单的小测验，结合具体实际业务，如图 5-8 所示是某公司 A 业务 1 月份每天的用户访问、点击、转化、交易、售后数据，首先，我们能否发现异常？

大家不妨拼接“数据敏感度”，用 1 分钟的时间从上述 31 天的 5 个指标数据中观察异常点，按照一列一列的顺序观察下能否看到异常数据点，算作一个简单的小测验，看看自己的数据敏感度能达到怎样的水平。

我相信任何人盯着上面的报表一直看都会晕得慌，其实我们也有替代的方案，那就是图表。我们不妨先看看第一列数据：

| 日期 | 访问量 | 点击量 | 临时订单 | 交易订单 | 售后咨询 |
|---|---|---|---|---|---|
| 1月1日 | 3199 | 2796 | 654 | 66 | 7 |
| 1月2日 | 3629 | 3137 | 456 | 50 | 9 |
| 1月3日 | 4269 | 3240 | 734 | 66 | 21 |
| 1月4日 | 4227 | 3578 | 544 | 76 | 17 |
| 1月5日 | 4234 | 3634 | 1100 | 100 | 13 |
| 1月6日 | 4212 | 3674 | 929 | 94 | 15 |
| 1月7日 | 5933 | 4865 | 936 | 94 | 10 |
| 1月8日 | 3805 | 4943 | 905 | 99 | 21 |
| 1月9日 | 4779 | 4413 | 832 | 87 | 5 |
| 1月10日 | 4370 | 3788 | 1145 | 83 | 19 |
| 1月11日 | 4231 | 4219 | 1087 | 156 | 29 |
| 1月12日 | 5181 | 4422 | 980 | 123 | 23 |
| 1月13日 | 5070 | 4160 | 1487 | 142 | 29 |
| 1月14日 | 5952 | 5046 | 1223 | 146 | 31 |
| 1月15日 | 5358 | 4857 | 1121 | 130 | 29 |
| 1月16日 | 5393 | 4642 | 1128 | 113 | 25 |
| 1月17日 | 5237 | 4144 | 916 | 207 | 31 |
| 1月18日 | 6212 | 5006 | 2568 | 294 | 25 |
| 1月19日 | 5613 | 4703 | 1828 | 317 | 28 |
| 1月20日 | 5693 | 4768 | 1547 | 280 | 35 |
| 1月21日 | 6889 | 5534 | 1508 | 248 | 47 |
| 1月22日 | 7005 | 6009 | 1597 | 272 | 34 |
| 1月23日 | 6005 | 5028 | 1082 | 186 | 34 |
| 1月24日 | 5952 | 4845 | 1378 | 195 | 41 |
| 1月25日 | 7042 | 5635 | 1834 | 233 | 51 |
| 1月26日 | 6808 | 5377 | 1761 | 243 | 47 |
| 1月27日 | 7175 | 5664 | 1872 | 253 | 46 |
| 1月28日 | 7700 | 6197 | 1875 | 247 | 55 |
| 1月29日 | 7632 | 6115 | 2831 | 295 | 64 |
| 1月30日 | 5843 | 4343 | 1461 | 95 | 55 |
| 1月31日 | 7541 | 6065 | 2178 | 260 | 53 |

图 5-8

如图 5-9 所示，我们可以相对明显地发现 1 月 7 日的数据异常增高，1 月 30 日的数据异常降低。


图 5-9

接下来第一个问题就是：

1月7日是不是异常增高？1月30日是不是异常降低？会不会只是随机波动？

在理解上面问题之前大家可以同样思考下：

1月8日算不算异常降低？1月11日算不算异常降低？

1月14日算不算异常增高？1月21日算不算异常增高？

到这里大家应该明白了我所说的“异常”数值，为什么称之为“异常”而不是“随机波动”？

判断依据一方面可以来源于你的数据敏感度，依照此项指标的历史波动范围，这个波动远超历史波动范围了。另一方面，如果我们详细深究的话可以使用六西格玛理论，计算这项指标历史正常波动的平均值 μ 和标准差 σ，如果当天的数值波动超过了 μ±3σ，即可视为当天数值为异常数值。

下面我们再介绍一个神奇的统计学工具：控制图与控制线。控制图与控制线设计的起源是由美国贝尔电话实验室（Bell Telephone Laboratory）质量课题研究小组过程控制组学术领导人休哈特博士提出的不合格品率 P 控制图。随着控制图的诞生，控制图就一直成为科学管理的一个重要工具，某些特别方面成了一个不可或缺的管理工具。它是一种有控制界限的图，用来区分引起的原因是偶然的还是系统的，可以提供系统原因存在的资讯，从而判断生产过程受控状态。控制图按其用途可分为两类，一类是供分析用的控制图，用来控制生产过程中有关质量特性值的变化情况，看工序是否处于稳定受控状；另一类，主要用于发现生产过程是否出现了异常情况，以预防产生不合格品。运用控制图的目的之一，就是通过观察控制图上产品质量特性值的分布状况，分析和判断生产过程是否发生了异常，一旦发现异常就要及时采取必要的措施加以消除，使生产过程恢复稳定状态。也可以应用控制图来使生产过程达到统计控制的状态。产品质量特性值的分布是一种统计分布，因此，绘制控制图

需要应用概率论的相关理论和知识。

简单来说，控制图主要是在工业领域用来判定产品的误差是否在正常的范围内随机波动还是系统设备出现了故障的一种判断方法，我们不妨把这种思路转移嫁接到日常数据监控中。我们知道每天的各类业务的各项指标都会随着时间的推移而波动，那么我们怎么确定这些波动是在正常范围内的波动还是异常波动呢？控制图与控制线可以帮助我们解决这样一个问题。

控制图的种类有很多，其实我们可以结合之前说的 6 西格玛理论去理解控制图，控制图就是给数据波动一个上限和下限，我们甚至可以直接用 μ+ 3σ 作为上限，μ- 3σ 作为下限，制作如图 5-10 所示。

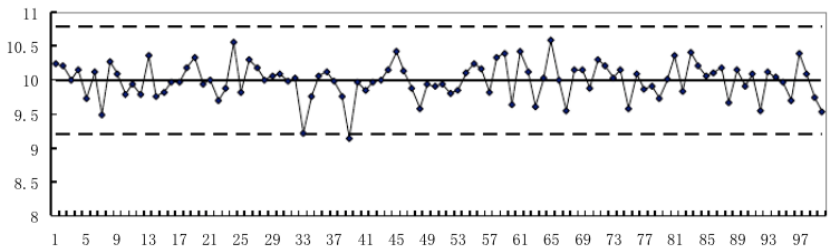



图 5-10

如果数据波动超出控制线，则可以视为异常波动。

当我们发现异常值之后接下来该做什么呢？

标红加粗放到邮件里？

查询到原因标红加粗写到邮件里？

给出解决问题的方案写到邮件里？

……

不同的人可能有不同的选择，但是大家可以看到的是越往后需要做的事情越多，发现问题往往只是异常数据监控的第一步。你发

现一个问题之后去追寻其中的原因是第二步，针对这个原因给出解决方案是第三步，拿出解决方案之后推动执行是第四步，执行之后看到数据效果是第五步，反思总结才算作数据异常监控工作的完结（图 5-11）。

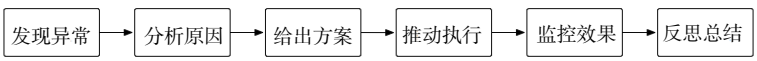



图 5-11

工作量是不是比想象中要多许多？为了进一步加深对异常数据分析的理解，下面用一个真实的异常数据分析案例来进行简单说明。

公司的数据分析师小 A 在 1 月 31 日观测 1 月 30 日交易量时，发现 1 月 30 日的数据异常低，小 A 第一件事情是通过邮件反馈了这个事情让大家周知，同时开始着手分析数据异常的原因。小 A 首先问技术部门昨天有无系统异常导致了用户无法使用相关功能，得到的反馈是一切正常。小 A 接下来找到市场部昨天有无进行相关市场活动，市场部反馈昨天同事把一个持续很久的活动下线了，因为市场部判断这个活动已经没人关注了。经过分析后小 A 判断此活动的点击率一直很高，不存在“缺乏关注的情况”。小 A 总结后发邮件阐述原因并给出方案。

A 计划：再次上线该活动

优势：时效性高，可立即执行

不足：活动持续时间较长，的确存在吸引力不足的问题

B 计划：再出一个类似的活动，维持活跃

优势：满足用户需求，创造新的刺激点

不足：需要设计周期，短期内无法上线

综合建议：A 计划先执行，做好 B 计划准备。

方案发出后，领导认同此方案并回复邮件：立即执行。

半小时后，小 A 询问市场部同事是否已完成上线？市场部同事

表示已经上线，正在着手准备新的替代活动。次日，小 A 到公司再次查询昨日异常指标是否已经恢复正常，确认数据已经回归正常水平之后发邮件告知相关同事：数据已经恢复正常。

最后，小 A 对此次异常数据的发现及处理做了总结：

（1）异常点被及时发现并且在第一时间查到原因、给出方案、付诸执行，值得表扬。

（2）市场部同事在没有事先查询数据与求证的情况下凭借主观判断下线相关活动存在操作风险，希望市场部负责人出台与活动上下线有关的管理规范，避免此类事情再次发生。

我们梳理下小 A 对整件事情的处理过程，如图 5-12 所示。

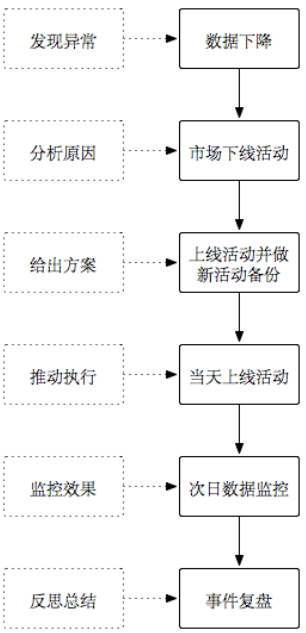



图 5-12

我们看到小 A 不仅发现了问题，解决了问题，而且在此基础上思考造成这件事情的原因是什么，提出了市场部规范相关管理制度

的建议。至此，整个异常数据分析的流程才算结束。这时我们再回顾整个异常数据处理的流程，数据异常分析是不是比想象中要复杂得多？亲爱的小伙伴们，你们掌握了多少呢？

## 5.4 MySQL 查询语言

介绍完报表制作与异常数据分析，大家对数据分析师有了一个粗浅的认识，接下来，要进行的 MySQL 是数据分析师的基础工具之一，掌握 MySQL 之后，我们在进行数据分析时才能做到有数据可分析。我们做数据分析对数据的时效性要求极高，如果每次都向技术部门提数据需求会极大降低数据分析的时效性。同时，由于数据分析的不确定性，有时候一天可能需要查询上百次数据，这么高频次的即时数据查询几乎不可能依托于别人的工作。与此同时，在掌握了 MySQL 查询语言之后，就可以进阶到数据分析的下一步：满足数据需求。接下来让我们花半个小时的时间，了解下 MySQL 查询语言，相信我，半小时足够了！

在进行 MySQL 学习之前我想再和大家确认两个与数据库紧密相关的问题：

- 数据是如何生成的？

公司的业务依托于互联网，当用户访问公司的业务时无论是通过网页端交互还是通过客户端交互，用户的每一次点击、每一次跳转都好比在交互页的这张纸上写字，用户在此留下了写字的时间、行为、内容，甚至可以记录用户的地理位置。同时在交互页可以设置按钮和功能，每当用户点击和操作时，我们把它称之为数据请求，我们会记录这些数据请求。这样就产生了庞大的数据。

- 数据是如何存储的？

就像大家知道的那样，数据存储在数据库里，数据库存储在服务器上。详细点说，每当我们记录一个用户的一系列行为时我们会

在数据库中生成一条记录，如表 5-2 所示。

表 5-2

| user_id | amount | type | create_time |
|---|---|---|---|
| 1 | 300 | 1 | 2016/1/9 |

当我们有很多记录的时候，就生成了一张表，如表 5-3 所示。

表 5-3

| user_id | amount | type | create_time |
|---|---|---|---|
| 1 | 300 | 1 | 2016/1/9 |
| 3 | 500 | 3 | 2016/2/5 |
| 5 | 900 | 1 | 2016/3/1 |
| 7 | 400 | 2 | 2016/2/5 |

当我们有很多表的时候，就生成了一个数据库，如图 5-13 所示。

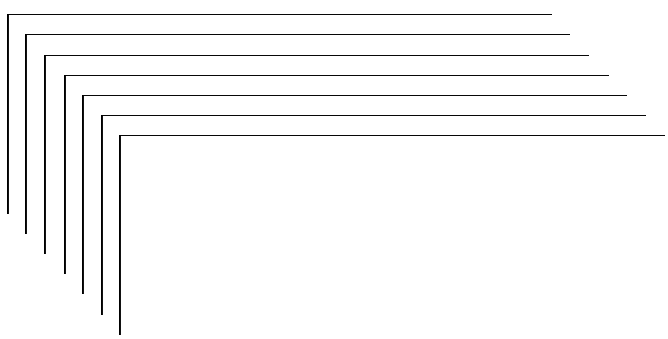


图 5-13

至此，我想大家对数据库应该有了一个直观的概念，接下来开始进入 MySQL 的学习。

首先花三分钟理清楚思路。

（1）SQL 语句的基本结构就是：

```
select a,b,c,d,e
from tableA
```

解释为：从 tableA 这张表格中选择 a,b,c,d,e 这五个字段（表格的表头）的所有记录（一行一行的数值）。

（2）如果不想选择所有记录，则需要加上限制条件：

```
select a,b,c,d,e
from tableA
where a>10
```

解释为：从 tableA 这张表格中选择满足 a>10 这个条件的 a,b,c,d,e。

（3）这时候你想对部分字段进行汇总求和，需要用到两个简单的函数 count() 计数和 sum() 求和：

```
select a,count(b),sum(c)
from tableA
where a>10
group by a
```

解释为：按照 a 为分类标准，看一看不同的 a 对应的 b 有几个，c 的总和是多少（类似 Excel 数据透视表）。

上面这些是不是很好理解？到这里我们已经大致了解了MySQL的基本结构。接下来介绍MySQL查询的核心功能：LEFT JOIN。

在进行数据库数据查询时，许多时候一张表无法满足需求，假设 TableA 是用户的交易信息（表 5-4），TableB 是用户的个人信息（表 5-5），我想知道在 A 表中交易用户对应的用户资料。

表 5-4（TableA）

| user_id | amount | type | create_time |
|---|---|---|---|
| 1 | 300 | 1 | 2016/1/9 |
| 3 | 500 | 3 | 2016/2/5 |
| 5 | 900 | 1 | 2016/3/1 |
| 7 | 400 | 2 | 2016/2/5 |

表 5-5（TableB）

| user_id | sex | age | active |
|---|---|---|---|
| 1 | F | 23 | 6 |
| 2 | F | 22 | 5 |
| 3 | M | 23 | 6 |
| 4 | F | 25 | 3 |
| 5 | M | 24 | 1 |

我们想要知道 A 表中交易用户对应的 B 表中的属性是怎样的，使用 Left Join，如下：

```
select *
from TableA a
left join TableB b on a.user_id=b.user_id
```

这段语句的意思是把 B 表中 user_id 与 A 表中 user_id 相同的项选出来拼接在 A 表的右侧，达到如表 5-6 所示效果。

表 5-6

| user_id | amount | type | create_time | user_id | sex | age | active |
|---|---|---|---|---|---|---|---|
| 1 | 300 | 1 | 2016/1/9 | 1 | F | 23 | 6 |
| 3 | 500 | 3 | 2016/2/5 | 3 | M | 23 | 6 |
| 5 | 900 | 1 | 2016/3/1 | 5 | M | 24 | 1 |
| 7 | 400 | 2 | 2016/2/5 | | | | |

首先，解释 TableA 后面的“a”，在进行数据拼接时，MySQL 要求每一张表必须有对应的表名称，我们把 TableA 命名为“a”表，同样的，TableB 后面的“b”是我们把 TableB 命名为“b”表，“on”后面接的是拼接条件，要求 a 表中的 user_id 与 b 表中的 user_id 相同。而 left join 的意思是左连接，即以“=”左面的 a 表为准， b 表中不存在 user_id 为 7 的记录，因此在进行左连接时此处为空值。如果我们把 left join 改成 right join：

```
select *
from TableA a
right join TableB b on a.user_id=b.user_id
```

将会出现如表 5-7 所示结果。

表 5-7

| user_id | amount | type | create_time | user_id | sex | age | active |
|---|---|---|---|---|---|---|---|
| 1 | 300 | 1 | 2016/1/9 | 1 | F | 23 | 6 |
| | | | | 2 | F | 22 | 5 |
| 3 | 500 | 3 | 2016/2/5 | 3 | M | 23 | 6 |
| | | | | 4 | F | 25 | 3 |
| 5 | 900 | 1 | 2016/3/1 | 5 | M | 24 | 1 |

最终结果将以“b”表的记录数为准。

再有 inner join 为取两者交集，full join 为取两者并集，此处不继续做展开。还有一个十分重要的功能就是，在多表选择的时候 select 后面的字段一定要带上表名，诸如：

```
select a.user_id,a.amount,b.sex,b.age
from TableA a
left join TableB b on a.user_id=b.user_id
```

其中“a.”月“b.”表示后面的字段来源于 a 表还是 b 表。

再加上筛选条件 where，这个语句就完整了。

```
select a.user_id,a.amount,b.sex,b.age
from TableA a
left join TableB b on a.user_id=b.user_id
where a.user_id>=3
```

得到结果如表 5-8 所示。

表 5-8

| user_id | amount | sex | age |
|---|---|---|---|
| 1 | 300 | F | 23 |
| 3 | 500 | M | 23 |

如果需要分组求和男性与女性的数据表现，只需要使用 group by ：

```
    select b.sex,count(a.user_id) user_num,sum(a.amount)
amount
    from TableA a
    left join TableB b on a.user_id=b.user_id
    group by b.sex
```

得到结果如表 5-9 所示。

表 5-9

| sex | user_num | amount |
|---|---|---|
| F | 1 | 300 |
| M | 2 | 1400 |

掌握了这些，读者基本就可以简单地查询数据库了，在日常工作中还会经常使用一些其他函数：left()、date()、count(distinct )、datediff()……当你需要使用的时候网上查询相关功能就可以，MySQL 查询语句是不是很简单？

## 5.5 数据需求处理

通过简单的学习和了解，我想大家已经对 MySQL 有了一定的了解和掌握，掌握了 MySQL 之后，我们对数据结构有一个相对清晰的认识和把控，接下来我们进入数据分析师日常工作的第三项：数据需求处理。

随着“用数据说话”的理念逐渐深入人心，公司各项业务在进行决策时需要依据各类数据来进行分析评估，这时大家的数据需求就会以各种形式出现在数据分析师的工作单上。理论上只要能够想到的就一定能够实现，通过对 MySQL 的灵活运用，所有的数据需求都能够被满足。然而，随着时间的积累，你会发现越来越多的现象：

- 需求已经是本周第三次提了，每次都是要更新数据。
- 需求提了改，改了又提，返工好几次。
- 需求很奇怪，看起来完全没什么用，只是需求方一念之间的突发奇想。
- 需求涉及太多公司的核心运营数据，需求方只是个外围推广人员

……

简单的数据需求却有太多的问题困扰着你，占用着太多的时间。跨部门沟通的效率低下是任何一个公司都存在的问题。这时我们需要一些数据需求规范，加一些数据沟通与处理的技巧。然而规范流程的问题在增加了规范制度的同时增加了沟通成本，降低了公司的自由度，一开始我们是非常自由的流程（图 5-14）。

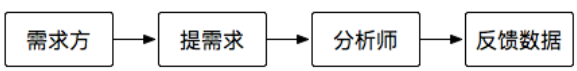



图 5-14

开始我们认识到许多核心数据不能直接面向所有人，于是我们加入审核流程（图 5-15）。

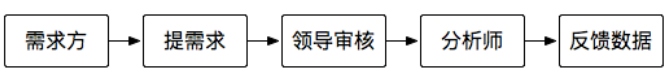



图 5-15

随着数据需求越来越多，我们开始让不同的分析师负责不同的模块（图 5-16）。

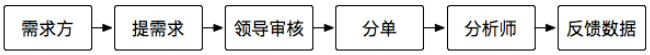



图 5-16

接下来，我们发现许多需求由于需求方不够明确，反馈的数据往往达不到目的，于是增加了二次沟通环节（图 5-17）。

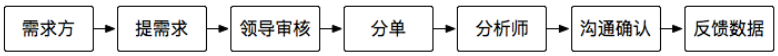



图 5-17

后来，发现许多人提了一些突发奇想无意义的数据需求，我们开始对数据的效果进行追踪验证，因此增加了后期追踪环节（图 5-18）。

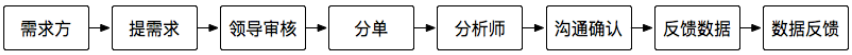



图 5-18

……

就像大家预料的那样，许多时候我们都是遇到问题解决问题，而当问题越来越多，解决方案越来越多的时候，整个过程会变得非常臃肿，我曾经尝试的能够解决完已知所有问题的流程图如图 5-19 所示。

大家明显地能够看到一个简单的数据需求经过这样的流程一折腾往往耗时又耗力，虽然能规避很多问题，但是许多可以“便宜从事”的事情都被规范了，在规避一切操作风险和不安全因素之后，严重影响了得到结果的我们得效率，这时再从总体上看这样一个问题：怎样为公司的业务决策提供数据支持？

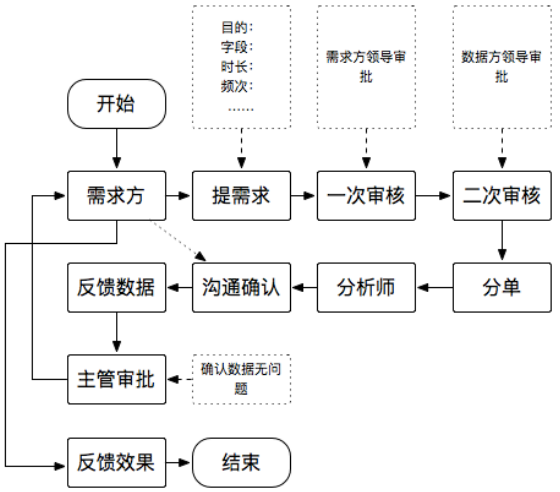



图 5-19

再次尝试回顾数据需求流程的更新与迭代，我们最关心的问题是什么？

**数据支撑决策**

那么你的决策是什么？或者说你的目的是什么？我的这些数据能够满足你的目的吗？你提的需求是正确的需求吗？从一个专业的数据分析师的角度应该怎样用数据实现你的目的？

思维的转变其实是化被动为主动，从被动的接收数据需求到主动地承担数据需求。在被动接收数据需求的时候总是会面临业务人员不懂数据的问题，沟通之间存在误差，数据人员不知道业务人员需要的东西真正是什么，所以在数据提供上就存在误差。同时，由于目的的不清晰导致认知偏差使数据人员不信任业务人员，业务人员无法完全依靠数据来作支撑。为了解决这一问题，我们在处理数据需求时第一件事情就是问：

**你的目的是什么？**

明确目的这一点非常重要，任何公司在实际运营中都是以结果为导向的，而你想要实现的这个目的即为结果，如果大家认可这一

个目的，为了统一的目标就能达到步调一致。同时当目标明确之后数据的效果便更好查验，当数据需求被满足后，需求方使用了这些数据是否达到了目的？如果达到了目的，那么是怎么达到目的的？如果没有，问题是什么？能否优化改进后进一步实现这个目的？

以结果和目的为导向之后，处理数据需求对于数据分析师就有了更多的要求，是否足够了解业务很多时候是数据分析师能否完成任务的关键。它要求数据分析师对业务员想要达成的目标进行数据拆分，把它变成数据可支撑的内容，然后再变为 MySQL 查询语句从数据库中导出，然后以最合理的方式呈献给需求方，让需求方以此来实现目标。在这样的过程中数据分析师把原本是需求方提的“数据需求”变成了自己给自己提“数据需求”，然后满足这样一个数据需求。这样做充分地利用了数据分析师更了解数据的优势，把业务员不懂数据的漏洞补全。依照这样一个流程，我们重新做了一个数据需求流程（图 5-20）。

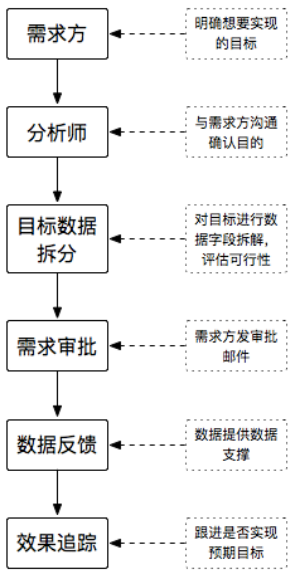



图 5-20

依照此数据流程，我们将数据需求的过程简化的同时也规避了

操作风险，同时保证了数据的准确性和时效性。为了便于理解，不妨以一个数据需求实例来看一下完整的数据需求是怎么完成的。

公司近期上线了新产品，运营人员小 A 想要知道市场反映怎样。于是在第一时间，他想了解使用新产品用户的以下几个特征：

性别分布、年龄分布、地域分布、设备类型分布

开始，数据分析师根据小 A 的需求做出了以下的数据报告，如图 5-21 所示。

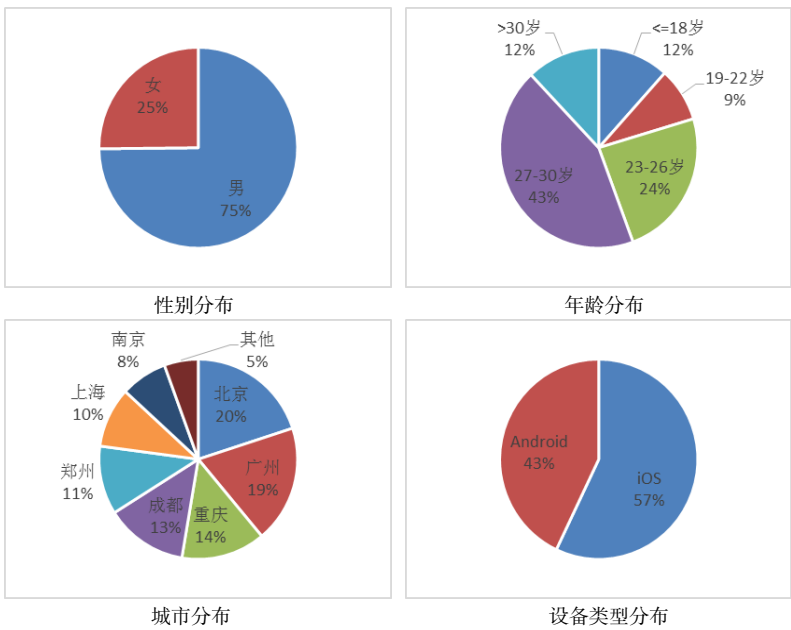



图 5-21

小 A 拿到这些数据之后做了总结：我们的新产品主要被 27~30 岁北京和广州地区使用 iPhone 手机的男性使用。这个总结可以说足够精简直白，基本涵盖了之前的数据需求所呈现的信息，那么这个数据需求就做完了吗？

我们在做一件事情的时候会评估做这样一件事情的投入产出比，即 ROI（Return On Investment）。假设投入了一个小时的数据查

询、调取、整理的时间得到了这样一个结果，那么这样一个结果有什么价值和回报呢？小 A 得到“我们的新产品主要被 27~30 岁北京和广州地区使用 iPhone 手机的男性使用”这样一个结论之后能用来做什么呢？难道只是“哦！原来如此！”满足了好奇心和兴趣吗？我们不妨去深究一下小 A 提这个数据需求的出发点和原始目的是什么。

数据分析人员找到小 A：“你提这个需求的目的是什么？”

小 A：“想做个用户画像。”

这就是最终目的吗？我们再来继续深究：

“为什么要做用户画像？”

小 A：“想做市场投放，想找一个用户特征进行突破。”

到这里目的就清晰很多了，我们再问：

“用户的哪些特征能够帮助你决策如何进行市场投放呢？”

小 A：“这我就不清楚了，一般都是从年龄、性别、地区等开始的。”

到这里我们就看到问题的所在了，作为需求方在一开始可能只是一个想法，在深究之后才明确他的目的是想通过用户画像来进行市场投放。关于最终目的的确认总是可以通过这样的问答来进行到最后，得到数据支撑必要的答案。如果我们继续问下去：

“为什么要进行市场投放？”

小 A：“近期用户增长面临瓶颈期，交易量增长放缓，利润增长跟不上了。”

直到我们问出“利润”这两个字，问答算是彻底完成了。无论对于任何企业，利润永远是最终的目的，一切的行为工作如果最终不能导向利润，那么这个行为就是没必要的，在这样的事情上花费时

间和精力就是对人力成本的一种浪费。毫不夸张地说，许多人文关怀和企业文化的目的是提高员工的状态增加产出，最终核心还是以利润为导向。

既然我们的分析师知道小 A 的目的是对用户进行画像，寻找用户特征以便于进行市场投放，那么我们的数据需求要从哪些维度来进行呢？或者说需要描述哪些字段呢？

任何对象的数据刻画都可以从这两个维度进行开展：属性、行为。

属性指的是用户自有的属性，不以其个人意志为转移。比如说一个人的年龄、性别、肤色、星座、出生地、居住地等，这些描述一个人的客观状态的信息为这个人的属性信息。行为指的是这个人进行了注册、提交资料、点击某页面、把某物品加入购物车、购买某物品等，这些记录一个用户在平台上进行种种操作的行为信息。小 A 想要对用户进行画像，寻找用户特征以便进行市场投放，我们将用户属性数据与行为数据进行结合，得到如图 5-22 所示的维度。

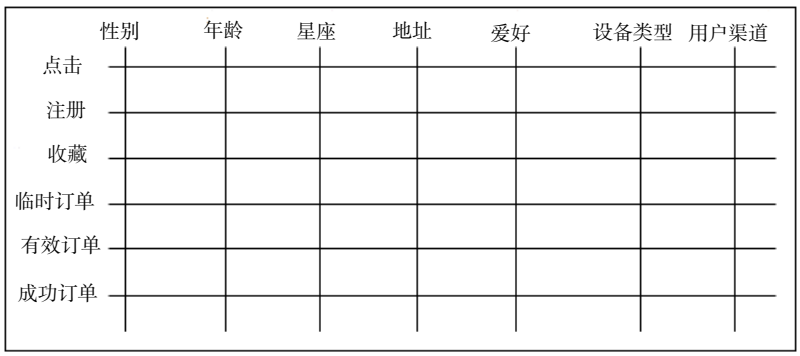



图 5-22

如果横向看上面的表格，用户的每一个行为都对应着用户在这个行为上的属性分布：

点击用户的年龄、性别、星座等分布情况

注册用户的年龄、性别、星座等分布情况

成功交易用户的年龄、性别、星座等分布情况

……

同样，纵向地看上面的表格，每一个属性的用户都对应着不同的行为表现：

男/女性在各个环节的数据表现

不同年龄段用户在各个环节的数据表现

不同渠道用户在各个环节的数据表现

……

如果我们详细地对每一个维度进行数据统计，那么上图中每两条线的交叉点都是一个数据需求，如果把这些需求全部完成的话无论是人力还是耗时都将达到一个不可承受的量级。这时则需要考量一下哪些维度是对实现目标“市场投放”有用的，我们要针对这些维度对用户进行数据调取和数据统计。

考虑到市场投放主要是以线上的形式，不太区分用户性别和年龄，反而是对地区和用户的渠道具有针对性，我们尝试对地区和渠道两个属性维度对应的用户行为数据进行数据统计。

首先对不同城市的用户行为进行数据统计得到如表 5-10 所示结果。

表 5-10

| 城　市 | 点　击 | 注　册 | 收　藏 | 临时订单 | 有效订单 | 成功订单 |
|---|---|---|---|---|---|---|
| 北京 | 357 | 214 | 107 | 86 | 43 | 39 |
| 广州 | 339 | 203 | 102 | 81 | 57 | 51 |
| 重庆 | 245 | 221 | 110 | 88 | 44 | 40 |

续表

| 城　市 | 点　击 | 注　册 | 收　藏 | 临时订单 | 有效订单 | 成功订单 |
|---|---|---|---|---|---|---|
| 成都 | 239 | 143 | 72 | 57 | 29 | 26 |
| 郑州 | 198 | 119 | 59 | 48 | 24 | 21 |
| 上海 | 174 | 104 | 52 | 42 | 21 | 19 |
| 南京 | 136 | 82 | 41 | 33 | 16 | 15 |
| 其他 | 99 | 59 | 30 | 24 | 12 | 11 |

制作折线图如图 5-23 所示。

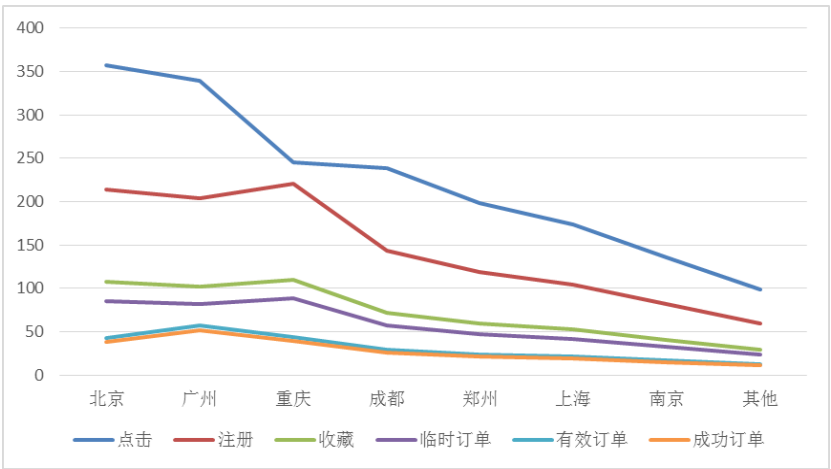



图 5-23

结合图表我们可以得到如下结论。

（1）北京、广州、重庆、成都四个城市的活跃度明显高于其他城市。

（2）同时广州地区临时订单到有效订单的转化率明显高于其他城市。

（3）重庆地区点击到注册的转化率明显高于其他城市。

接下来我们对不同渠道的用户行为进行数据统计得到如表 5-11 所示结果。

表 5-11

| 渠　道 | 点　击 | 注　册 | 收　藏 | 临时订单 | 有效订单 | 成功订单 |
|---|---|---|---|---|---|---|
| 渠道 A | 487 | 292 | 146 | 117 | 58 | 53 |
| 渠道 B | 402 | 241 | 121 | 96 | 68 | 61 |
| 渠道 C | 316 | 190 | 95 | 76 | 38 | 34 |
| 自然传播 | 231 | 139 | 69 | 55 | 28 | 25 |

制作折线图如图 5-24 所示。

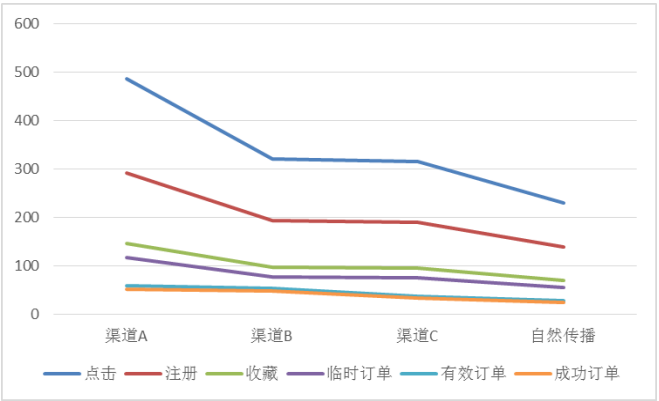



图 5-24

同样，我们能够得到渠道 A 的获客明显高于其他渠道，渠道 B 的临时订单到有效订单的转化率高于其他渠道。

通过这两份数据的调取，小 A 可以清晰地看到接下来要在北京、广州、重庆、成都四个城市的渠道 A 和渠道 B 进行重点投放，同时借鉴广州地区临时订单到有效订单转化的方法，重庆地区点击到注册转化的方法，将渠道 B 的临时订单高转化率复制到渠道 A 上。这个时候我们回归之前第一次得到的结论：“我们的新产品主要被 27~30 岁北京和广州地区使用 iPhone 手机的男性使用”，哪一个结论

更具有指导意义呢？结论不言而喻了。

现在我们已基本了解数据需求的处理方法了。听起来简单的数据需求，处理起来可一点都不简单。做数据需求时要明白许多时候需求方提的需求并不一定是真实想要的数据，而分析师要纠正这样的错误，分析需求方真正需要的数据。你要的不一定是你想要的，我给你的才是你真正想要的。在这样的层面上大家能更好地理解数据分析师其实是业务人员而不是技术人员。

最后一点小提议：数据分析师一定要有自己的需求表，并且做一个优先级排序，许多时候我们会埋头于数据中晕头转向，一个简单的需求列表配上优先级，会让你对数据工作的把控更加得心应手。

## 5.6 进行项目分析

在了解了数据需求的处理方法之后，下面要一起了解数据分析师的核心工作：项目分析。在讲述项目分析怎么做之前我们来明确一些简单的概念：什么是项目？某百科对项目的介绍：是指一系列独特的、复杂的并相互关联的活动，这些活动有着一个明确的目标或目的，必须在特定的时间、预算、资源限定内，依据规范完成。概念理解起来还是有一些抽象，如果我们把项目与任务做个比较就好理解很多了，不妨举个实例。

任务：小A，今天晚上把宿舍卫生打扫了。

项目：小A，制订一个轮换打扫卫生的制度。

任务往往是单一的执行与服从命令，而项目往往是一系列任务的合集。同时项目具有鲜明的目的性，承接项目的人一定是清晰地了解项目的目的是什么才能承接项目。项目中有较强的主管性和能动性。而任务更倾向于简单的执行，强调服从和执行力。

自然而然地，项目经理作为项目的负责人需要制订项目的实现方案与执行方案。项目经理（Project Manager）是现在互联网公司

普遍存在的产品经理的前身，只是现在互联网行业的产品经理更倾向于对产品的逻辑和用户体验负责。我们这里说的项目经理更倾向于广义上的项目经理。为了实现一个目标而建立的项目，需要项目经理从项目发起之后进行方案策划、方案执行、问题处理、流程把控、结果验收等全流程的开展工作。高级数据分析师很多时候要充当项目经理的角色，从项目的发起开始运用数据的力量驱动整个项目的执行与落地。

数据分析师的日常会遇到哪些项目呢?

我们可以先简单地把项目分为被动的项目与主动的项目，前者是被动指派完成的项目，后者是自己主动探索摸索的项目。一般数据分析师在一定阶段都会有一个自己的探索性分析项目，同时日常开展被动的项目分析。被动的项目又可以细分为优化类项目和拓展类项目，优化类主要指的是在公司已有的业务和流程之上进行优化，这类项目可以理解为在别人修建好的框架上进行二次梳理。拓展类项目则多为公司想要投资或者投放一些资源去开展新的业务，需要数据分析师提供项目数据分析用来评估预期效果，最终来评估项目的 ROI，确认是否有必要开展此类活动。主动项目主要是依据公司的实际业务，诸如用户分析、商户分析、A 业务前瞻性分析、K 地的市场潜力分析等。这些项目主要以探索为主，结果往往导向一个新的商机或者发现公司业务潜在的问题，主观性较多，如图 5-25 所示。

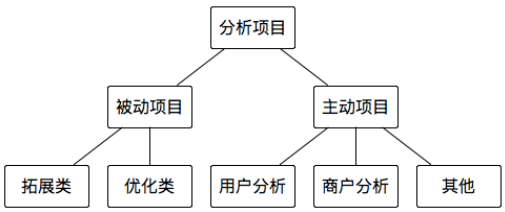



图 5-25

不同类型的项目分析会在日常分析工作中不断地出现和重复，

对被动项目的完成水平决定了分析师能否胜任这份职位，对主动项目的探索成果决定了分析师的上升空间。被动项目一般都有明确的时间要求与效果要求，而主动项目的分析一般无明确的要求，只能看最后的商业效果。

了解了数据分析项目的内容和构成之后，要如何完成一个数据分析项目？我相信有许多经验丰富的项目经理，处理过很多问题，任何一个项目安排下来都可以直接上手开干，凭借丰富的经历和阅历，遇到问题解决问题，能把事情做得很漂亮。还有一类人喜欢先搭建框架，再搭建主干、明确细节，然后落地执行，最后遇到问题解决问题。前者一力降十会，后者运筹帷幄决胜千里。哪一种是更好的选择呢？或者说哪一种方式更适合数据分析的项目呢？

为了便于大家理解，我们把这两种解决问题的方式简化为如图5-28所示的两组流程。

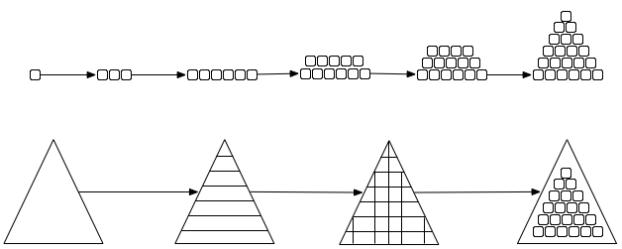


图 5-26

同样为了搭建一个金字塔，第一种处理问题的方法倾向于一个一个累积，从底层开始搭建，逐层完成，第二种处理问题的方法倾向于先搭建一个大框架，然后把框架里的主要结构层次梳理清晰，把较细的网络梳理完成，最后再向这个完整的架构里填充内容。两种不同的工作方式，但是明显第二种方式更加理性和科学。第一种方案在实施的过程中许多不可知的问题都是在实际执行中才会发现，之前的一个环节疏忽了就可能导致整个项目前功尽弃，而第二种方案则更加合理与可控，可能出现的问题已经被具象化和模块化，

遇到问题只需要解决相应的模块就好，不会涉及其他模块，在方法上比第一种方案更胜一筹。

互联网时代的发展极其迅速，许多项目都是尝试甚至是创新，不可能有极其丰富的历史经验来辅助决策，这时我们能否保持清晰的思路就很重要，我们在做前期的准备工作时把框架结构梳理得十分清晰再下手执行会极大地提高效率和增强对项目的把控力。

下面不妨用一个数据分析项目实例来展示项目性数据分析是如何开展的。

（以下数值仅供参考案例，不代表实际学校情况）

公司将要在 A 学校组织一场活动，下面是活动的计划表，如表 5-12 所示。

表 5-12

| 项　　目 | 具体内容 |
| --- | --- |
| 活动名称 | 春天来啦 |
| 负责人 | 大 A |
| 参与人 | 小 B、小 C、小 D、小 E |
| 时间 | 2016 年 2 月 28 日 |
| 地点 | 北京×××大学 M 校区 K 食堂旁 |
| 目的 | 获客 |
| 内容 | 发放小礼品邀请参与活动 |
| 预算 | 6000 元 |

在没有数据分析师的情况下这个项目的可执行性建立在主管在对学校的基本认知和本月财务预算是否充足上，然后在一些主观因素的参与下判定这个项目能否执行。当我们的数据分析师介入这个项目时会有怎样的处理方式呢？一个最简单的问题：评估这件事情的投入产出比，给出预期效果的数据分析。

数据分析师要想评估该活动的预期效果，基本的逻辑就是：活动能够覆盖到多少人？这些人群中能够有多少人会对产品感兴趣，对产品感兴趣的同学又有多少能转化到试用，最后有多少同学愿意为这件事情花钱，我们能够从这里挣得到少钱。

经过分析师的调查：该学校的 M 校区在校生有 20000 人，选取的 K 食堂位于四个宿舍的附近位置，四个宿舍大约有 5000 名学生，活动从 10 点开始至 19 点结束，覆盖了学生午餐和晚餐的时间。由于活动时间为周二，80%以上的学生都会上课，也即会在食堂附近就餐。与此同时，由于活动宣传，会把校园内分布在别处的学生吸引过来， 预计转化其他学生中的 10%。综上此次活动大约能够覆盖 5000×0.8+15000×0.1=5500 人。根据历史活动学生群体转化率经验，覆盖到转化的用户大约 15%，因此，当天愿意尝试产品的学生人数大约为 825 人。

分析到这里，数据分析师有两种选择：一是继续计算下去，看一看会有多少用户会为此付钱，在未来的一段周期内能否用利润覆盖现在的获客成本；另一种是比较之前的其他活动，比较单一获客成本，是否好于平均水平。第一种方案更加科学，第二种方案更加高效，这时我们就要具体问题具体分析了。鉴于目前互联网公司多以融资铺垫前期成本为主，未来的利润来源有许多渠道，当下多以获客为主，我们多使用第二种方案来进行评估项目的合理性。我们接着看上面的项目，成本是 6000 元，大约转化学生群体 825 人，获客成本=6000/825=7.27 元，计算数值表格如表 5-13 所示。

表 5-13

| 项　　目 | 具体内容 |
| --- | --- |
| 在校人数 | 20000 人 |
| 一级传播用户 | 5000 人 |
| 二级传播用户 | 15000 人 |
| 覆盖用户 | 5500 人 |

续表

| 项　　目 | 具体内容 |
| --- | --- |
| 转化用户 | 825 人 |
| 投入成本/元 | 6000 元 |
| 获客成本/人 | 7.27 元 |

计算获客成本之后，下面要评估获客成本的高低。一般情况下会将此项目的获客成本与历史平均获客成本进行比较。如果之前计算过平均获客成本为 P，分析项目就可以截止到这里了，只需比较 7.27 与 P 的大小就可以做一个简单的判断。但是如果我们之前没有计算过平均获客成本，这里有一个简单的计算方式。我想大多数公司都有邀请制，给用户现金或优惠券作为补贴，我们只要计算这个邀请制补贴的提现率再乘以贴现成本就可以计算获客成本。至此，这个市场投放项目的数据分析就算结束了，活动执行层面则需要项目负责人具体开展。

这个被动类的拓展性数据分析项目中，数据分析师更倾向于扮演一个项目助理的角色，方案计划好后执行层由项目负责人来完成。而项目性分析中被动类的优化类数据分析就需要分析师直接化身项目经理来执行了。

这里再举一个被动类的优化类数据分析项目。

今年经济不景气，公司老板准备减少市场投放的同时提高用户转化率，要求数据分析师分析现在的用户转化率，出一个方案，把转化率提升 1~2 个点。这个项目的难度要明显高于上一个项目，项目只有一个目标：提升转化率，但是如何寻找提升的点，怎么做方案以及如何执行都是未知数，在这种情况下我们该如何着手呢？

这时不妨看看上文提到的处理项目的两种方法（图 5-26）。

如果我们把最后实现转化率的提升比作把金字塔搭建完成，我们同样有两种方案。

第一种方案，如图 5-27 所示。

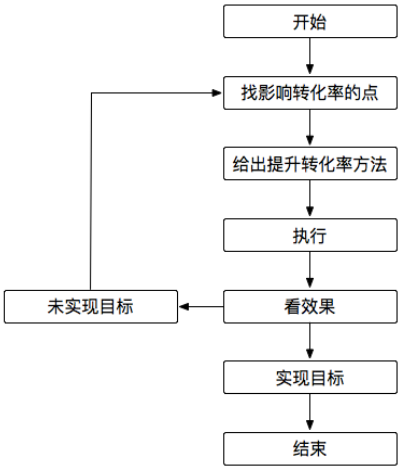



图 5-27

第二种方案，如图 5-28 所示。

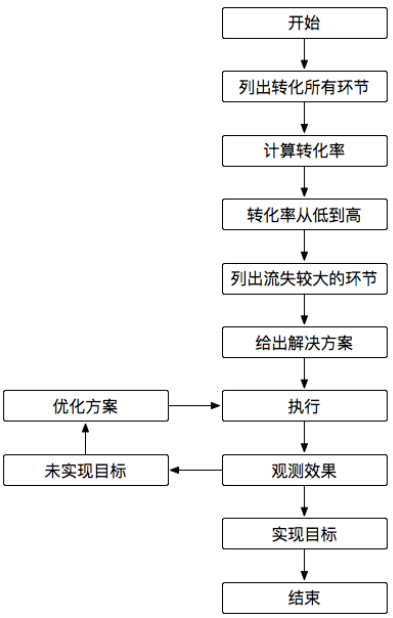



图 5-28

大家观察这两种方案的差别在哪里？第一种处理问题的方式是线性的，实现不了目标就再找一个点重复一遍流程。而第二种解决问题的方式则是框架式的解决问题，自上而下，先列出解决问题的范畴，再在这个范畴里进行循环与流程处理。第二种方法显然优于第一种方法，这个项目的实际处理方法先留做思考，我们会在之后章节的数据分析实战中具体讨论。

到这里，我想大家对数据分析中的最重要的一个环节项目性分析的流程已经有了相对具体的认识。接下来我们来谈谈项目性分析中的核心：制订方案。

制订方案是把分析结果落地执行的核心环节。任何项目在梳理完思路找到原因之后，要将头脑中的想法落地实现则需要制订方案来帮助实现。简单的方案诸如上文中你发现校园推广活动的人均获客成本显著高于历史均值，你给的方案是校园推广直接不做了还是要让市场部再设计一个方案？停止校园投放活动是你的方案，让别人制订方案也是你给的方案，哪种方案才是合理的？若是复杂一点的项目，例如你发现了提高用户转化率的核心是用户在第一次注册时就尝试使用公司的产品，这部分用户的后续转化率都很高。那你应该制订怎样的方案才能让用户第一次注册就尝试使用公司的产品？加大补贴力度？提高推广人员的整体素质？还是在用户引导环节投入更多的精力进行产品优化？这些都是方案，哪一个方案才是合理的、最优的、可执行的？

制订方案的关键思路是：找到问题原因，穷举所有可能方案，然后比较每个方案的优劣，之后判断筛选方案的可行性，最后给出最优解，如图 5-29 所示。

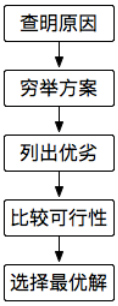



图 5-29

这五个环节看起来很简单，但是每一个环节后面对应的工作量都是及其巨大的。我们还是用上文提升转化率的那个项目来分析。

假设我们发现用户转化率下降的最大环节是用户注册后不再使用我们的产品，我们需要一个方案来改善这一现状。

第一步我们需要做的是查明原因，从题目中看到影响用户转化率低的原因是用户注册后不使用我们的产品。首先对转化率进行拆解：

转化率=使用用户数/注册用户数

从上面的公式中很容易看到想要提高转化率有两种途径：增加使用用户数或者是减少注册用户数。显然，通过减少注册用户数来提升转化率的手段必然导致业务量萎缩，这样即使转化率提升了但业务量却下降了，可以说得不偿失，我们只能用提高使用用户数这样的方法。如此，我们的问题就变成了：

为什么用户注册了却不使用产品？

我们需要寻找原因，怎么寻找原因呢？从已有的数据中寻找原因。但是问题是我们没有或者只拥有很少的未使用产品的用户的信息，往往只有一个手机号，我们有的绝大多数信息都是来源于注册了又使用了产品的用户信息。这个时候不妨换一个角度思考问题：如果知道用户为什么使用产品，找到这些元素然后去满足那些未使

用产品的人，问题不就解决了吗？所以我们把问题变成了：

为什么用户注册后使用我们的产品？

这个问题就有迹可循了，毕竟我们有庞大的历史用户数据库。如何分析这些用户呢？大家还记得之前说过从用户属性和用户行为两个角度进行用户分析吗（图 5-22）？

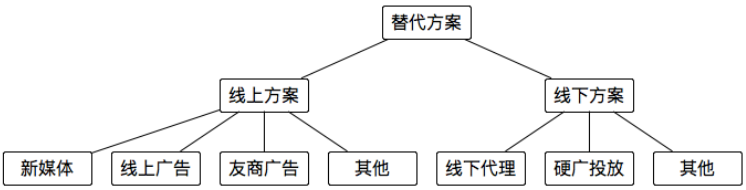
同样，我们需要对已有的成功交易的用户从上面这些维度进行试探性分析，只是我们的目的从以前的市场投放变成了现在的用户消费行为分析。

分析的过程我们就不再重复了，假设最终得到的结果是：此类用户注册的主要原因是渠道代理商给的补贴，用户注册拿到补贴就离开了，留存率较低。发现原因之后我们开始给出解决方案的第二步：

穷举所有可能的方案。

在发现这个问题后我们有两种选择：一是放弃这个渠道；一是优化这个渠道。除此之外就没有别的可能了。放弃渠道这个选择，需要停止合作然后寻找替代的渠道，当然不寻找替代渠道直接放弃也是一种选择，但是这种选择不太实际，我们只考虑寻找替代渠道的方案。如何寻找替代渠道呢？不妨把替代渠道先分成线上替代和线下替代。线上替代可以分为：新媒体替代、线上媒体投放替代、友商合作替代方案等。线下替代可分为线下代理人替代、硬广投放等替代方案。

我们首先把暂停合作寻求替代渠道的所有可能穷举一遍如图 5-30 所示，做这件事情我们可能需要市场部的同事协助一起完成，毕竟数据分析人员相较于市场部同事，后者更加了解渠道，数据分析师在这个层次扮演的角色更倾向于引导者和项目梳理者。




5-30

接下来，我们需要对按类分好的渠道进行优劣分析与对比，在每个渠道下面列出该渠道对应的优势与劣势。这样我们对暂停合作寻求替代渠道的方案分析完毕，通过对每个渠道的优劣对比可以挑选出最理想的几个渠道作为替代渠道，截止到这里第一类方案梳理完毕。接下来讲优化已有渠道的方案，我们优化渠道有两个方案：一个方向是优化自身的方案、策略和补贴制度；另一个方向是优化渠道代理的工作方式与推广形式。优化自身的方案可以细分为降低目前的补贴减少羊毛党的数量，提高获得补贴的门槛等。优化渠道代理可以选择可信的代理人或是用公司员工替代，新增渠道代理管理制度，设置用户活跃百分比要求等。

同样，我们把可以优化的方案穷举，如图 5-31 所示，可以邀请产品同事一起探讨有哪些优化的方向和优化点。在穷举完这些方案和框架之后同样需要列出每一个模块的优劣，然后进行比对，最终选出优点最突出而劣势很少的优化方案。

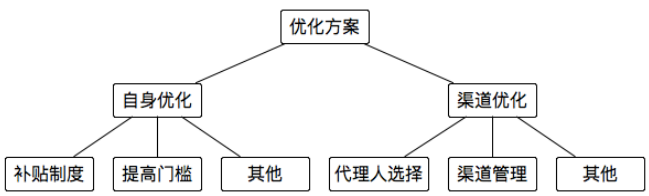



图 5-31

在进行上面的工作之后。我们就可以把挑选出来的不同方向的方案列出来，给出每个方案的优势与劣势，这就是方案制作的第三个环节，这个环节和穷举方案结合得很紧密，在穷举所有可能的同

时就可以进行评估。方案优劣的评估对数据分析师业务能力要求很高，要求数据分析师能掌握每个业务的核心关键点，用数据分析的思维对他们进行拆解和划分，最终做出选择。

做完上面的工作，我们已基本完成制订方案的工作，最后一步就是评估所挑选方案的可行性了。可行性评估包含了这个方案落地执行需要多长时间、多少人力、多少物力、能产生多大的效果等。这些工作大多需要跨部门进行沟通，是时候组织一个项目会议了，把相关部门负责人聚在一起评估方案的可行性，数据分析师统筹规划会议流程、把控会议进度、总结大家的意见给出最终结论。至此，方案已经基本完成，最后，数据分析师将成型的方案发送大家二次确认，经反馈无误后就可以付诸执行了，至此我们的方案制作环节就结束了。

方案制作作为项目性分析的核心环节大家应该形成自己的想法，同样，项目性分析这个数据分析师的核心工作至此也基本完结。最后，我们不妨再次强调一下项目性分析是一个庞大的工程，切忌心浮气躁急于求成，眼高手低是一切新手都会犯的错误，希望大家都能抱有强烈的责任意识，遇到问题解决问题，犯了错误就想办法纠正错误，在项目性分析的道路上越走越远。

## 5.7 数据分析的结构化梳理

通过上面六节的介绍我相信大家已基本了解数据分析师的主要工作内容了，数据分析的四大模块，每一个模块都对应很多内容需要了解和掌握，总体上，我们再回顾梳理数据分析师的基本工作（图 5-32）。

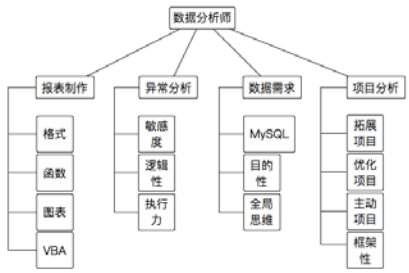



图 5-32

数据分析的这四大模块希望大家能熟练掌握。随着经验的积累和对业务的把控力，大家对数据分析的认识会越来越多，会遇到自己的瓶颈期也会有自己的迷茫期，渐渐地，你会发现数据分析这份工作有越来越多的知识需要掌握，到那时才能称为一名合格的数据分析师。

# 第6章 数据分析师进阶

公元前 5 世纪，芝诺发表了著名的阿基里斯悖论：他提出让乌龟在阿基里斯前面 1000 米处开始，和阿基里斯赛跑，并且假定阿基里斯的速度是乌龟的 10 倍。当比赛开始后，若阿基里斯跑了 1000 米，设所用的时间为 *t*，此时乌龟便领先他 100 米；当阿基里斯跑完下一个 100 米时，他所用的时间为 *t*/10，乌龟仍然前于他 10 米。当阿基里斯跑完下一个 10 米时，他所用的时间为 *t*/100，乌龟仍然前于他 1 米……芝诺认为，阿基里斯能够继续逼近乌龟，但绝不可能追上它。

一尺之棰，日取其半，万世不竭。

## 6.1 思维与态度

通过上一章的学习大家对数据分析师的工作已基本掌握，然而，普通数据分析师与高级数据分析师的差异有一个非常重要的点：数据思维。数据思维与数据敏感度有一些类似，都是类似于情商类的看不见摸不着的东西。简单来说数据思维是一种通过数据手段解决问题的思维。大家还记得中学时期或是大学时期的数学证明题吗？

已知条件 A、B、C、D 条件，要求证明 E 是成立的。

一道证明题往往只是一句话，然而解题过程往往要占据一整页篇幅。几何证明题出现的频次更是尤其高，还记得我们在进行数学证明的时候做的证明流程吗？几乎所有的证明题都是要求通过已知条件转换为未知条件，而我们证明的过程恰恰是方向解剖，如果要 E 成立需要什么条件？假设需要 E、F 成立；E、F 成立有需要 G、H、I 成立；G、H、I 成立恰好需要 A、B、C、D 条件，证明完毕。

证明流程如图 6-1 所示。

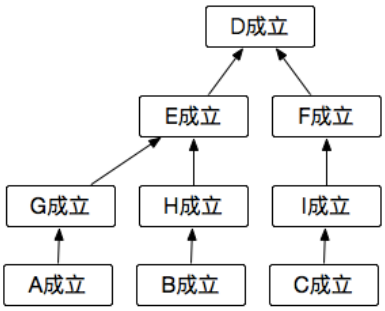



图 6-1

其实这就是一种以结果为导向的思维方法，数学带给我们的思维最重要的体现就是在解决问题的方式上。证明题的流程之所以如此清晰严谨多是因为出题者已经事先梳理了证明逻辑，对于解题者来说正确答案只有一个：证明 D 成立。

除了证明题，我们还经常面对的另一类问题是应用题。应用题大多是把日常生活场景抽象简化，在题目中描绘一个场景，常见的题型可以归类如下：

小明在×××的时候发现，A 事件有 a 属性，B 事件的值是 b，假设小明的 C 属性数据是 c，问小明在 D 时的值 d 是多少？

这类题目刻画了一个事件场景，大多会交代时间、地点、人物、事件，然后给出一些参数，要求另外一个参数的值。同样，我们想要知道 D 的值需要两个条件 E、F，想要知道 E、F 的值需要条件 G、H、I，而 G、H、I 的值可以通过 A、B、C 的值 a、b、c 求得。逻辑关系梳理完成后需要通过对 a、b、c 三个数值进行加减乘除简单的数学计算或是积分求导等高阶数学算法，最终求得结果 d。应用题和证明题的区别在于它在证明题的逻辑思维基础之上增加了数值运算。

随着应用场景的不断复杂，我们引入了一元一次方程、二元一次方程组、黎曼积分、极限思想等这些数学工具。这些工具发明的初衷在于解决实际生活中遇到的问题，只是实际生活中遇到的问题

被抽象成了应用数学题。数学工具的不断丰富和复杂，人们不再拘泥于现实的应用场景，开始把数学研究单独作为一门技能进行拓展和延伸。于是产生了另一类数学题。

已知公式 A，条件是 B，当 n 趋向于正无穷，求 D。

A 是 B 的全覆盖，求证：C 是 D 的全覆盖。

P(A|B)=K，求 P(C|A)。

……

此类问题已经是进阶到高等数学的范畴了，高等数学与普通数学的最大区别就在于其应用场景没那么明确具体，不像加减乘除能够让你买菜，高等数学更加抽象和理论化。它们对应的是极限的思想，全面拆分问题的思想，这时我们再看看本章开头的两个实例：

公元前 5 世纪，芝诺发表了著名的阿基里斯悖论：他提出让乌龟在阿基里斯前面 1000 米处开始，和阿基里斯赛跑，并且假定阿基里斯的速度是乌龟的 10 倍。当比赛开始后，若阿基里斯跑了 1000 米，设所用的时间为 $t$，此时乌龟便领先他 100 米；当阿基里斯跑完下一个 100 米时，他所用的时间为 $t/10$，乌龟仍然前于他 10 米。当阿基里斯跑完下一个 10 米时，他所用的时间为 $t/100$，乌龟仍然前于他 1 米…… 芝诺认为，阿基里斯能够继续逼近乌龟，但绝不可能追上它。

一尺之棰，日取其半，万世不竭。

这是极限思维的实际案例，大家有没有发现问题在哪里呢？留作课后思考题吧！想清楚了自然豁然开朗，想不清楚可以去找能够帮助你想清楚的方法，寻找答案的过程也算是数据分析思维的一部分。

我们看到上文给出的数学问题的三个模块其实对应着数学思想的变化，如图 6-2 所示。


图 6-2

数学从提供解决问题的方法到变成数学工具，再变成数学思想。这一演变的过程为我们提供了解决问题的思路，思考问题的方法。数据分析的思维可以借鉴数学思想的内容，从解决实际问题的角度出发，找到需要解决这个问题的元素，一层一层地剥离下去，最终联系到我们已有的资源。同样，我们抛开数据分析的实际应用场景去探索数据分析方法的优化空间和可行性，对已有的数据进行聚类、分类等探索性分析，提升数据的使用效率，挖掘数据中潜在的价值，这些就是数据分析的思维方式。

数据分析的思维是一种解决问题的方式，以结果为导向的向数据源头的追溯。数据分析师要有一种遇到问题解决问题的自信。没有问题是无法解决的，没解决的原因只能是投入大于产出，解决该问题带来的收益小于投入。

技能是容易掌握的，但是思维却是很难培养的。从我们接触数学这门学科的那一天开始，数学就尝试向我们传递这样一种思维方式，因此，在面试数据分析师时我往往会问一问面试者的数学成绩怎样。数学成绩能够部分反映一个人对数学思维的理解与运用，即使他自己都可能没有意识到这一点。这些关于数学解题的思维方式正是数据分析师所需要的，也是数据分析师必备的。那么，如何培养数据分析的思维呢？不妨先培养解决数学问题的思维。经常做一些逻辑推理题或是看一些侦探小说，会有帮助的。

数据分析思维一方面体现在它的逻辑性和方向性，另一个重要特征是绝对客观与绝对理性。“不以物喜，不以己悲”的态度对于数

据分析思维来说很重要，它能够帮助你摒弃主观的偏见与看法。诸如遇到突发事件能在第一时间冷静下来，抛去恐慌的情绪；对自己喜欢的项目客观分析，不对数据进行修饰；对自己犯下的错误能客观评论，给出解决方法等。喜怒哀乐是每个人都会有的情绪，而对数据分析师而言，一旦进入工作就要绝对理性与客观，这也是数据分析师思考问题的前提。

任何人都会犯错误，我们在日常工作中难免会犯错误，作为数据分析师，每天都和一大堆数据打交道，稍有不慎就会犯错误。如何对待自己犯下的错误是衡量一个数据分析师处理问题客观性的重要标准。人们在面临指责时的本能反应是逃避或是反击，这是人性的弱点，数据分析师能否克服这样的弱点将是他能否进阶的重要因素。当领导指责你工作没做好的时候你会以怎样的态度去面对这个问题？

攻击的态度：不是我的错，是什么什么原因造成的。

逃避的态度：好像是错了，对不起！

客观理性的态度：是我错了，纠正方法是×××2 小时内可以完成。此次错误的原因是××××，以后不会再犯了，本月绩效相应的部分会进行扣除。

如果你是领导，你会喜欢哪种态度呢？

领导永远是以结果为导向的，指责你犯错或是沉浸在内疚的情绪中于事无补，第一时间应该做的事情是把结果做好，然后再进行自我检讨，用最客观的态度进行自我批评。这样不仅给自己一个教训，也会让领导不会因此过度责怪你。你已经给出了面对此错误的最好的解决方案，别人也不会再节外生枝。更大的可能是领导会因为这件事增加对你的好感度与信任度。

我想大家都读过历史类或是战争类的小说，谋士给统帅的策略一般会给出上策、中策、下策，而统帅经常会出于人道主义原则选

择中策或是下策。越是厉害的谋士给出的策略出发点越是绝对理性，不考虑感性的情怀与仁慈，一切以成功为最终目的。高阶的数据分析师就要具有这种谋士的精神，客观与理性的解决问题。同样，只要统帅提出问题，谋士总能给出解决方案，虽然有些理想主义的情怀，但是能从一定意义上反映数据分析思维的两个方面：分析问题的思想；处理问题时的态度。

思维与态度作为数据分析思维的两个核心要素是衡量一个数据分析师水平的软指标，培养自己的思维与处理问题的态度需要在实践中不断完善和进步。“学而不思则罔，思而不学则殆”，数据分析的过程需要大家不断思考、不断实践，才能在这样一个过程中不断提升自己。

## 6.2 软件升级：R or Python

随着我们数据分析能力的不断提升，Excel 渐渐无法满足日常需求，我们需要更专业化的软件来帮助做数据统计和数据分析。相应的问题就来了：统计学软件那么多？该选择哪一个？目前市场上的统计学软件包含但不限于：SPSS、R、Python、SAS、JMP……目前市场上较为火热的两款软件一个是 R，另一个是 Python，就像许多教程里描述的那样，“选择 R 还是 Python，这是一个问题！”

这两个软件之所以能够取得如此多的关注，部分原因是来源于大家对其他同类软件的不接受。SPSS 的操作可谓是傻瓜级的，点点鼠标就好了，对编程的要求很弱，与多数人眼中的高阶软件有些出入，于是就这样被忽视了。我想 SPSS 的发明者一定没有想过自己的软件会因为过于简洁易上手而被忽略吧！SAS 软件是出了名的难以安装，在软件安装上就能将一大半的初学者拦在门外。SAS 高达 8 个 G 的内存占有量，配合着高昂的价格几乎不适用于个人数据分析师，即使软件破解也需要花费很大的精力，破解失败就可能面临重装系统的尴尬局面。其他的统计软件相对小众，这样一来 R 与

Python 就因为它们的容易安装、易于上手和编程自由度高的特性脱颖而出，于是引出了问题：R or Python?

R 作为开源的免费软件，很多时候无论是更新速度还是自由度都相对较高。R 进行数据处理时的 data frame list 等数据处理方式使数据清洗变得极为便捷。R 很容易上手，查到基本的命令和包，直接“print”指令就有结果了。但是如果要自己写算法、优化性能的时候，学习难度陡增。R 语言在可视化方面的表现尤为出色，一些必备的可视化软件包 ggplot2、ggvis、googleVis、rCharts 非常强大。当然 R 语言有它的问题：诸如处理数据较慢同时不容易深度学习。Rstudio 非常不错，提供类 Matlab 环境。

R 语言的便捷性使统计人员处理问题更加轻松，但是相应的代价是电脑的运行速度可能很慢。虽然 R 的体验是缓慢的，但是有多个包来提高 R 性能：pqR、renjin、FastR、Riposte 等。另一方面 R 学习起来并不容易，特别是从 GUI（图形用户界面）来进行统计分析。如果你不熟悉它，那么即使发现包可能会非常耗时。

Python 作为开源软件的好处是可以做很多事情，包括做网页搭建数据框架，其作为数据分析模块 pandas 也十分专业。Python 作为一种通用的语言，容易和直观。在学习上会比较容易，它可以加快写一个程序的速度。此外，Python 测试框架是一个内置的，这样可以保证代码是可重复使用和可靠的。Python 把不同背景的人集合在一起。作为一种常见的、容易理解的，大部分程序员都懂的软件，可以很容易地和统计学家沟通，你可以使用一个简单的工具就把你每一个工作伙伴都整合起来。Python 绝大多数的帮助文档都比 R 好了许多。有些包用起来没有 R 方便。Windows 下有 python(x,y)，还有许多商业工具。

同样，Python 也有它的缺陷，一个很重要的问题就是数据的可视化。可视化是选择数据分析软件的一个重要标准。虽然 Python 有一些不错的可视化库，如 Seaborn、Bokeh 和 Pygal。但相比于 R，呈现的结果并不总是那么顺眼。

总的来说，Python 对于 R 来说是一个挑战者，它不提供必不可少的 R 包。虽然它在追赶，但是还不够。如果只是处理“小”数据，用 R。结果更可靠，速度可以接受，上手方便，多有现成的命令、程序可以用。要自己搞个算法、处理大数据、计算量大的，用 Python。开发效率高，一切尽在掌握。

大家尽可以从各种角度对这两款软件进行比对和分析，但是唯一需要和大家说明的是不要浪费太多的时间进行纠结，可以放心大胆地采用先入为主的策略上手学习。浪费太多的时间在纠结到底应该学习 Python 还是 R 上只会让自己犹豫不前。选择自己的目标，大胆上手吧！

## 6.3 数据分析师的格局

究竟是“屁股决定脑袋”还是“脑袋决定屁股”？

在讲完数据分析的思维与态度之后，我们再与大家探讨数据分析师的进阶之路。进阶有两种途径，一种是“屁股决定脑袋”，当你的职位上升之后需要承担更多的任务和把控更大的局面，你的思维自然而然地随着提升；另一种是“脑袋决定屁股”，当你的思维足够把控更大的局面时，你的职位会随着你思维的上升而上升。前者的进阶方式是被动的，后者的进阶方式是主动的。然而对于数据分析师来说，后者尤为重要。

数据分析师作为新兴的行业与传统职业的最大区别在于它的职业体系化不成熟，不像传统行业依据工作年限的积累，熬到一定资历和熟练度后升职加薪自然而然就来了。而现在互联网行业大潮的最大特色就是没有明确的等级制度和薪资制度，扁平化普遍存在于整个行业，数据分析作为新的行业尤其如此，此时“脑袋决定屁股”是必然趋势。“心有多大，舞台就有多大！”虽然是口号，但是在这里却尤其适用。下面我们就一起来看看数据分析师的格局。

我们的进阶过程如图 6-3 所示。

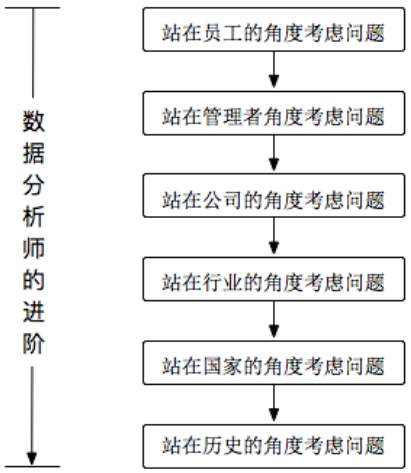



图 6-3

首先，我们先说站在员工的角度考虑问题。作为公司的一员你为什么工作？或者说你工作的目的是什么？

- 没工作就没钱，工作是为了生存。
- 我想学习这项技能，这份工作能够帮助我成长。
- 不工作又干什么呢？在家多无聊。
- 我想成就一番事业，我厌倦了枯燥无味的生活……

不同的人有不同的答案，作为数据分析师的你也有自己的答案。如果我们总结一下工作的目的，大体可以分为五类：

**工资、技能成长、行业资历、人脉关系、成就事业**

近一半的员工工作的目的是第一项：工资；还有三分之一的员工工作的目的是第二项：技能成长；剩下的六分之一可能是为了行业资历或是人脉关系又或是成就一番事业。那么作为数据分析师的你工作的目的是什么呢？由于数据分析师的技巧性和工作特质，大多数的工作时间都是面对一大堆报表进行演算推理，一定程度上导

致了许多人目光的局限性：我的这个分析方法对不对？该怎样优化这个算法？聚类分析的显著性能否得到保证？绝大多数的数据分析师把自己工作目的的第一位留给了技能成长，其次是工资。由于大数据概念的火爆，大家对数据分析师的未来充满了希望，同时由于数据分析行业不够成熟，大多数情况下我们都在摸索着前进。这也就造就了大家目前的格局。基于这样的目的我们会有怎样的期许呢？

- 早点发工资吧！
- 这个软件现在很热门，我要掌握它！

这是基于目前的主要目的产生的心理期许，对薪资的渴求和对成长的渴望。

这样有什么不对吗？都对！只是这样的心理预期会让人偏向于执行层，偏向于完成指定的任务和学习掌握新的技能。这是一种员工执行者心态，温顺而机警，服从命令听指挥的同时维持着自我学习的状态。其实维持着学习的状态就已经超过了大多数的普通求职者了，工作中不学习不求上进的大有人在。

接下来，我们站在管理者的角度考虑问题，即以结果为最终导向。当我们在接到一个任务时会关注很多问题，比如这个任务需要别的部门配合，数据库的底层框架不够优化等非个人原因导致这个任务无法完成，于是，作为数据分析师的你把这些困难按条列出反馈上去。这没什么不对，你的领导面对你提的这些问题有两个选择：一是放弃这个任务，因为困难太多；二是针对这些困难给出解决办法，然后再让你继续执行这个任务。你认为领导更倾向于做出哪一种选择呢？我想大多数人都会选择第二种：解决困难然后继续执行的方案。其实这里能否理解：以结果为导向这一条非常重要。领导只关注结果是否完成了，这个时候作为执行方的直接反馈却是有许多困难无法克服，然后让领导帮忙处理这些困难是典型的执行者思维。这些事情又不是我的问题，我为什么要管？那么这些问题是谁的问题呢？如果你把这些问题和困难的关注点转移到以结果为导向

的层面呢？

我为了实现一个目标，结果遇到了许多问题，于是我把这些问题全部解决了，然后把结果反馈上去。领导在乎的是结果，你把结果给他了，顺带把涉及的所有问题给解决了，你做的就不仅仅是执行者的工作了，你已经在这个层级上提升了一层，把执行的好坏转移到了结果的好坏，思维层级的提升带来的是思考问题方式的提升，当你能够站在这个角度考虑任务和解决问题时，你离晋升就不远了。

恭喜你从一个员工的思维方式提升到了一个领导者的思维方式。你在自己的职业生涯中迈出了关键性的一步。那么接下来我们继续探讨思考问题方式的下一个层级：站在公司的角度考虑问题。

公司的目的是什么？一定是盈利。这一点毋庸置疑！公司的目的不是解决就业问题，不是为了给员工提供福利，不是为了融资，不是为了上市……公司的终极目的一定是盈利，只是实现盈利这个目的的方式有所不同而已。当然有些非盈利机构另当别论，他们把“非盈利”三个字放进了组织类型里自然不在此列，那我们来谈谈公司的盈利问题。

- 你知道你干的活能为公司带来多少收益吗？
- 你知道你们部门能为公司带来多少收益吗？
- 你知道人事在节假日给员工发的福利能为公司带来怎样的收益吗？
- 你知道为什么有些公司安排员工出行一定是五星级酒店吗？
- ……

你有太多的不知道，所以你不是公司的 CEO，你只是个普通的员工或是领导者。如果我们站在公司的角度考虑盈利问题就能帮我们想明白许多事情，许多拿不准的决定就能拍脑袋，许多不明白的决策就很容易理解。公司的目的是盈利，那么怎么盈利呢？不同的公司有不同的盈利模式，但是盈利的数值来源于收入减去支出，如何控制支出增加收入，则需要我们做投入产出比控制。这么想来我

们在做决策的时候如果清晰地了解项目的投入产出比就可以很好做决策了。当然决策不仅是评估两个数值的大小，还会有许多标记效应，诸如对公司品牌、软实力等情况的影响，同时项目的成长性、周期性都会纳入决策中，而这些都需要计入我们的投入产出比模型。另一方面，做好决策和方案之后还需要员工进行执行，员工执行的质量效率决定了这个方案能否取得预期的效果。是超出预期还是没有达到预期，这里面包含了人员的把控、考核制度、公司执行力等要素。这时我们一般有两个维度去思考这个问题：一个是硬指标，比如通过合理的 KPI 绩效考核、科学的管理制度和沟通方法让员工认真对待工作；另一个是软指标，主要体现在员工的自我驱动力。我们为员工提供一个舒适的环境、美好的未来、向上的心态或是集体荣誉感。这些都是在构建公司的软实力，通过这种方式提升员工的产出，实现公司的唯一目标：盈利。

许多时候我们说站在领导者的角度考虑问题能够做一个好员工，站在 CEO 的角度考虑问题能够做一个好领导，那么如何做一个好的 CEO 呢？我们不妨看一看在 CEO 之上的层级是什么？茫茫多的 CEO 引领了他们的行业，许许多多的行业又构成了国家市场经济环境，不同国家经济体系相互交流冲突形成了国际贸易。国际贸易又是从大航海时代一直发展至今形成了历史……随着思考方式的层级不断提升，自身的认识难免会出现局限性，这里不再做过多展开，大家不妨信马由缰地去思考：

- 我们所处的行业处于一个怎样的成长期？
- 国家目前的经济政策对我们所处的行业有怎样的影响？
- 我们所处的行业对国家的经济有什么推动作用？
- 国家的经济的发展与别国经济发展的现状是怎样的？
- 全球经济环境对我国的影响是什么？
- 未来全球的经济将以怎样的形式发展？

……

这些问题大家不妨结合自身的情况想一想，开阔自己的眼界和学识。有这样一条永恒的经济学定律：风险越大，收益越高。你能够承担得起更大的风险也就能获得更大的收益，那么如何承担起更大的风险呢？开阔自己的视野，不断增加自己能够把控的事物。相信“心有多大，舞台就有多大”。

# 第7章 数据分析实战

目前，市场上仍有大量的公司没有自己的报表系统，大家对数据报表的认知仍旧停留在观察每日新增用户和交易用户这些核心结果数据上。数据的作用主要是用来了解事情的结果怎样，而对于为什么产生这样的结果却是全然没有数据支撑。

下面我们一起从零开始搭建这样一套体系。

## 7.1 报表系统

通过前六章的学习，我们对数据分析的工作已经基本熟悉，数据分析第一层次的工作就是报表制作，我们每天制作许许多多的报表，这些报表有的用来反映获客数据，有的用来反映转化数据，有的用来反馈交易数据，我们把这些报表放在一起就形成了报表系统。接下来让我们从零开始搭建一套报表系统吧！

一般来说，搭建一套报表系统有五个环节：梳理业务逻辑、技术数据埋点、报表结构、数据调取、报表系统（图 7-1）。

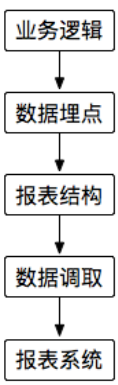



图 7-1

让我们首先从业务逻辑的梳理开始。

数据报表是用于为业务决策服务的，因此在构建数据报表之前

需要对业务逻辑进行梳理。不同的公司有不同的运营模式，对应的数据报表系统自然是各不相同。但是万变不离其宗，我们不妨以一家电商公司为例，把公司的业务简化为三大模块：获客、转化、交易。获客是通过各种推广手段使得用户注册或是下载成为公司用户的统称，转化是公司用户对各类产品产生意向的过程，交易则是用户为公司的产品付钱的过程。获客的途径可以有好多种，用户的自传播、通过竞价排名、第三方联合推广、地推团队、硬广……我们需要公司获客的每一个渠道方式，明确新增的每一个用户的来源。对用户转化的监控更倾向于是对用户流失的统计，用户在不同环节的转化率对应的就是用户在每一个环节的流失情况，用户产生交易意向的临时订单，用户决定要支付的有效订单，用户成功付款的成功订单，争取做到每一个意向用户是否进行到下一环节都能清晰明了。对交易的统计绝对是重中之重，这是公司盈利情况的核心要素，许多时候我们都要对公司交易数据进行保密。交易数据中包含了用户支付的金额、产品的原价、盈利与亏损情况，等等。

这样，就得到了数据报表体系的三大模块：获客模块（图 7-2）、转化模块（图 7-3）、交易模块（图 7-4），每一个模块都对应着细分的渠道、环节、流程，我们对每一个环节和流程的业务都要清晰熟悉，争取做到业务中每一个点都没有被遗漏。

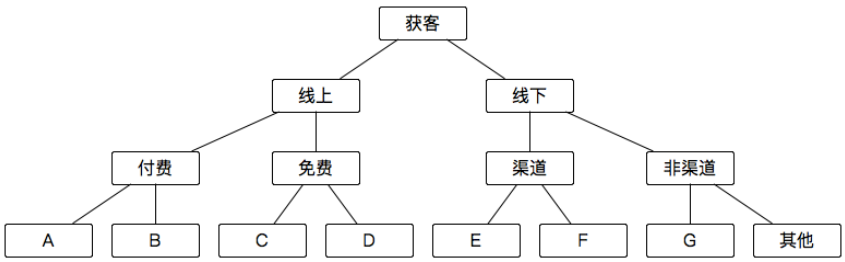



图 7-2

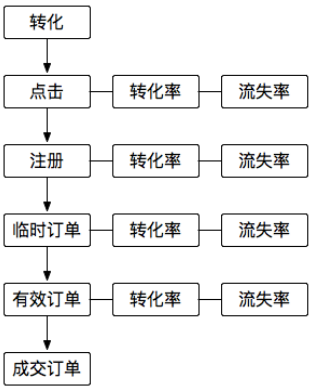



图 7-3

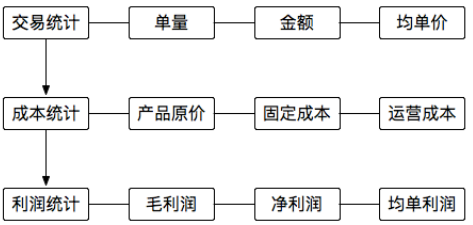



图 7-4

上面的环节需要结合公司的实际业务进行具体了解和分析，数据报表系统的搭建是基于业务的，所以对业务的了解程度在很大程度上决定了数据报表系统能否真正发挥监控业务、指导业务的作用。

梳理完业务之后面临的问题是：能否统计到这些数值？我们清楚地知道了自己需要在哪些环节进行监控，但是这些环节在数据库中是否有对应的数据统计呢？对于一个没有数据分析系统的公司来说，数据库的底层框架多是基于对业务功能的支撑，而诸如用户在哪个环节点击了什么多是被忽略的，这个时候就需要数据分析人员把自己的需求转化成数据库可识别可转化的内容。如果细分，还有数据产品经理这样一个岗位，专门是为数据埋点与数据可视化设置的岗位，不过体量较小的公司不会为此单独设置岗位。所以这里需要数据分析人员自己完成。

其实，基于我们对数据库的理解和业务知识的把控，数据埋点这项工作其实很简单。要做的无非是对不同渠道不同环节的用户进行单独标记。没有标记的给出标记码，没有统计的嵌入数据统计。放在数据结构里就是某一单独标识的用户在某一环节做了什么，如表 7-1 所示。

表 7-1

| 用户 ID | 渠道码 | 统计页面 | 行　　为 | 时　　间 |
|---|---|---|---|---|
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |

给出明确的标识之后交给技术人员就好了，他们会用自己擅长的手法对数据结构进行搭建和优化。

在业务梳理和数据埋点完成之后，我们就可以正式进行报表输出了。报表输出的形式多种多样，最基础的是数据分析人员每天通过 MySQL 查询语句进行人工统计和汇总，通过邮件的形式发送给数据需求方。进阶一点的是把数据报表做成标准化的格式，每天通过系统自动触发查询生成表格然后邮件自动发送。还有一种是搭建数据可视化后台，每个人有自己的账号和权限，登录之后即可查看数据。越往后人工介入得越少，但是对于数据分析师来说，从最基础的数据调取开始往往是最好的，对业务数据的了解和把控就是从查询数据制作报表开始的。用眼睛看永远不如用手写来得印象深刻，同样，直接读报表去看数据表现永远没有一边制作报表一边看数据报表来得深刻直观。

首先是获客表，简单的样表如图 7-5 所示。

| 日期 | 关注人数 | 下载人数 | 转化率1 | 注册人数 | 转化率2 | 渠道1 | 占比1 | 渠道2 | 占比2 | 渠道3 |
|---|---|---|---|---|---|---|---|---|---|---|
| 2016/1/1 | 986 | 854 | 87% | 800 | 94% | 314 | 39% | 254 | 32% | 157 |
| 2016/1/2 | 1273 | 1180 | 93% | 964 | 82% | 372 | 39% | 279 | 29% | 179 |
| 2016/1/3 | 1068 | 983 | 92% | 834 | 85% | 304 | 36% | 277 | 33% | 120 |
| 2016/1/4 | 1354 | 1084 | 80% | 916 | 85% | 356 | 39% | 263 | 29% | 151 |
| 2016/1/5 | 1493 | 1050 | 70% | 878 | 84% | 308 | 35% | 288 | 33% | 176 |
| 2016/1/6 | 1499 | 1220 | 81% | 967 | 79% | 381 | 39% | 294 | 30% | 185 |
| 2016/1/7 | 1111 | 1019 | 92% | 847 | 83% | 383 | 45% | 217 | 26% | 104 |
| 2016/1/8 | 1381 | 1028 | 74% | 854 | 83% | 380 | 44% | 228 | 27% | 110 |
| 2016/1/9 | 1529 | 1003 | 66% | 906 | 90% | 391 | 43% | 234 | 26% | 149 |
| 2016/1/10 | 1456 | 1069 | 73% | 927 | 87% | 384 | 41% | 280 | 30% | 166 |
| 2016/1/11 | 1523 | 960 | 63% | 910 | 95% | 385 | 42% | 270 | 30% | 160 |
| 2016/1/12 | 1496 | 920 | 61% | 893 | 97% | 348 | 39% | 268 | 30% | 199 |
| 2016/1/13 | 1261 | 1006 | 80% | 877 | 87% | 366 | 42% | 300 | 34% | 123 |
| 2016/1/14 | 1893 | 1141 | 60% | 962 | 84% | 333 | 35% | 295 | 31% | 184 |
| 2016/1/15 | 1348 | 995 | 74% | 812 | 82% | 330 | 41% | 274 | 34% | 111 |
| 2016/1/16 | 1563 | 1033 | 66% | 824 | 80% | 305 | 37% | 244 | 30% | 143 |
| 2016/1/17 | 1655 | 1047 | 63% | 894 | 85% | 377 | 42% | 249 | 28% | 127 |
| 2016/1/18 | 1058 | 764 | 72% | 725 | 95% | 312 | 43% | 234 | 32% | 129 |
| 2016/1/19 | 1611 | 1068 | 66% | 867 | 81% | 366 | 42% | 250 | 29% | 119 |
| 2016/1/20 | 1127 | 805 | 71% | 783 | 97% | 332 | 42% | 214 | 27% | 166 |
| 2016/1/21 | 1242 | 1047 | 84% | 963 | 92% | 383 | 40% | 277 | 29% | 174 |
| 2016/1/22 | 1235 | 894 | 72% | 743 | 83% | 308 | 41% | 278 | 37% | 100 |
| 2016/1/23 | 899 | 815 | 91% | 749 | 92% | 366 | 49% | 229 | 31% | 104 |
| 2016/1/24 | 1131 | 1017 | 90% | 810 | 80% | 348 | 43% | 209 | 26% | 136 |

图 7-5

这张表相对清晰地反映了每天公司的获客情况，逻辑还算清晰。但是有一个非常严重的问题是数字不够直观，这个时候我们需要图表来进行数据可视化。参照前面章节的讲解，我们这里不妨用折线图反映每天注册用户的增长情况，下面是基于注册用户数做的两张折线图（图 7-6 和图 7-7）。

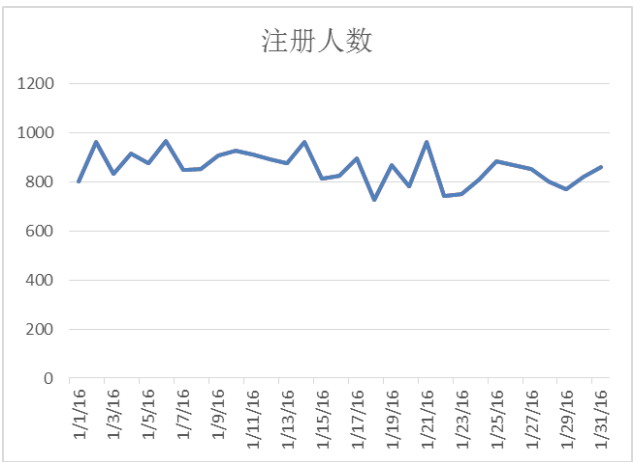



图 7-6

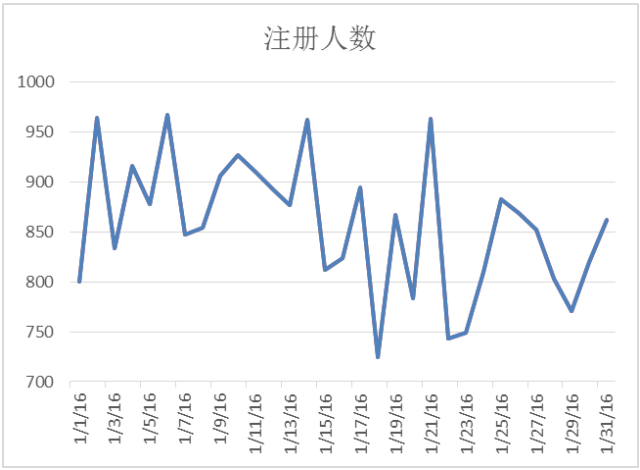



图 7-7

大家有没有注意到这两张图表的差异在哪里？

纵坐标的起点不一样！

一个是从 0 开始的纵坐标轴，一个是从 700 开始的纵坐标轴。相应地，我们在看这两张表时所能产生的结论可能就不一样了。看第一张图我们可能会说一月总体获客情况稳定，无明显波动。但是，看第二张图我们会发现每天新增用户的波动极大，同时月末呈下降趋势。哪一个结论是准确的？这就引出了一个问题：数据的欺骗性！或者是我们常说的“数据会撒谎”。我们在进行数据统计和数据可视化时一定要结合具体业务进行操作，保持客观性是进行图表制作的一个重要准则，大家不妨尝试选取不同的标准作图看看有什么差异？

我们统计 1 月不同渠道在获客上的占比，制作如下两张图表（图 7-8 和图 7-9）。


图 7-8

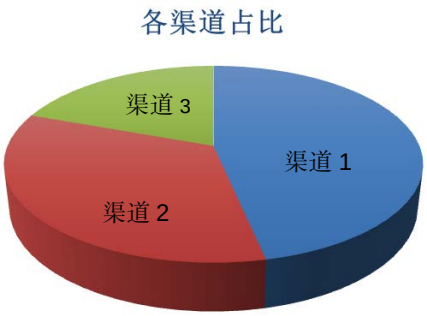



图 7-9

大家看这样两张饼图有什么不一样吗?

第一张图展示了渠道 1 大约占了获客总量的一半，渠道 2 和渠道 3 占了剩下的一大半，渠道 2 比渠道 3 多一些。但是第二张图直观地感受是渠道 1 和渠道 2 占了总体的 80%，渠道 3 占了剩下的一小部分。两张表因为作图的方式不一样，渠道 2 的获客量在第一张图是和渠道 3 相较，而在第二张图中就和渠道 1 相较了。通过简单的视觉变换，渠道 2 的地位上升了。所以我们在报表制作的过程中一方面要保证数据的准确性，另一方面要保证作图的客观性。

同样，当我们把目标转向转化率表（图 7-10）和交易表（图 7-11）时，大体上可以采用下面的方式对用户进行统计和监控。

| 日期 | 页面一 | 页面二 | 转化率 | 流失率 | 页面三 | 转化率 | 流失率 |
|---|---|---|---|---|---|---|---|
| 2016/1/1 | 986 | 854 | 87% | 13% | 800 | 94% | 6% |
| 2016/1/2 | 1273 | 1180 | 93% | 7% | 964 | 82% | 18% |
| 2016/1/3 | 1068 | 983 | 92% | 8% | 834 | 85% | 15% |
| 2016/1/4 | 1354 | 1084 | 80% | 20% | 916 | 85% | 15% |
| 2016/1/5 | 1493 | 1050 | 70% | 30% | 878 | 84% | 16% |
| 2016/1/6 | 1499 | 1220 | 81% | 19% | 967 | 79% | 21% |
| 2016/1/7 | 1111 | 1019 | 92% | 8% | 847 | 83% | 17% |
| 2016/1/8 | 1381 | 1028 | 74% | 26% | 854 | 83% | 17% |
| 2016/1/9 | 1529 | 1003 | 66% | 34% | 906 | 90% | 10% |
| 2016/1/10 | 1456 | 1069 | 73% | 27% | 927 | 87% | 13% |
| 2016/1/11 | 1523 | 960 | 63% | 37% | 910 | 95% | 5% |
| 2016/1/12 | 1496 | 920 | 61% | 39% | 893 | 97% | 3% |
| 2016/1/13 | 1261 | 1006 | 80% | 20% | 877 | 87% | 13% |
| 2016/1/14 | 1893 | 1141 | 60% | 40% | 962 | 84% | 16% |
| 2016/1/15 | 1348 | 995 | 74% | 26% | 812 | 82% | 18% |
| 2016/1/16 | 1563 | 1033 | 66% | 34% | 824 | 80% | 20% |
| 2016/1/17 | 1655 | 1047 | 63% | 37% | 894 | 85% | 15% |
| 2016/1/18 | 1058 | 764 | 72% | 28% | 725 | 95% | 5% |

图 7-10

| 日期 | 成功订单 | 成功金额 | 均单金额 | 成本 | 利润 | 均单利润 |
|---|---|---|---|---|---|---|
| 2016/1/1 | 223 | 481,222.58 | 2110.63 | 463,962.12 | 17260.46 | 75.70 |
| 2016/1/2 | 216 | 516,740.51 | 2117.79 | 494,022.48 | 22718.03 | 93.11 |
| 2016/1/3 | 216 | 531,051.85 | 2132.74 | 510,053.79 | 20998.05 | 84.33 |
| 2016/1/4 | 236 | 487,818.31 | 2139.55 | 468,799.87 | 19018.44 | 83.41 |
| 2016/1/5 | 213 | 500,555.39 | 2103.17 | 477,164.11 | 23391.28 | 98.28 |
| 2016/1/6 | 232 | 445,607.45 | 2111.88 | 432,218.46 | 13388.99 | 63.45 |
| 2016/1/7 | 239 | 480,988.75 | 2118.89 | 459,231.04 | 21757.71 | 95.85 |
| 2016/1/8 | 250 | 435,469.65 | 2124.24 | 418,055.57 | 17414.08 | 84.95 |
| 2016/1/9 | 246 | 477,670.71 | 2142.02 | 458,261.63 | 19409.08 | 87.04 |
| 2016/1/10 | 218 | 467,944.49 | 2146.53 | 450,140.61 | 17803.88 | 81.67 |
| 2016/1/11 | 223 | 476,761.82 | 2128.40 | 463,058.77 | 13703.05 | 61.17 |
| 2016/1/12 | 225 | 519,553.13 | 2138.08 | 498,529.28 | 21023.85 | 86.52 |
| 2016/1/13 | 217 | 519,164.75 | 2119.04 | 499,473.62 | 19691.12 | 80.37 |
| 2016/1/14 | 211 | 450,582.86 | 2125.39 | 435,208.35 | 15374.51 | 72.52 |
| 2016/1/15 | 203 | 504,141.12 | 2145.28 | 486,134.33 | 18006.79 | 76.62 |
| 2016/1/16 | 243 | 511,856.26 | 2132.73 | 498,593.89 | 13262.37 | 55.26 |
| 2016/1/17 | 239 | 453,479.50 | 2139.05 | 435,360.58 | 18118.92 | 85.47 |
| 2016/1/18 | 246 | 481,003.94 | 2109.67 | 463,025.64 | 17978.30 | 78.85 |
| 2016/1/19 | 248 | 463,883.25 | 2118.19 | 449,820.82 | 14062.43 | 64.21 |
| 2016/1/20 | 240 | 424,209.61 | 2100.05 | 404,225.43 | 19984.19 | 98.93 |

图 7-11

至此，我们完成了日常三大模块的报表统计，这些报表无论是人工手动查询还是做成系统自动触发，都是需要每天更新的报表。除了这些每天发送的日报，我们的报表系统里还需要加入周报、月报、季度报表等，这些概括性的报表是对周报的概括和总结。

在此基础之上，我们还会有用户留存率报表、用户活跃度报表、使用时长、使用频率、设备终端统计表等，用来反馈用户的数据表现。

首先，我们来看用户活跃度表，用户活跃度报表展示当天和 15 天用户的活跃成分，并提供活跃程度成分用来做进一步分析，帮您宏观了解用户的活跃程度及其活跃成分占比。活跃度统计一般会统计当日活跃成分和 15 日活跃成分。当日活跃成分展现每个天为单位的每个时间点的当日活跃用户的活跃程度。将当日活跃用户按照过去 15 天（含当天）启动的天数分为 1 至 15 组，计数并展示。

活跃 1 天的用户，表示这个用户在过去 15 天中仅有 1 天启动；

活跃 2 天的用户，表示这个用户在过去 15 天中仅有 2 天启动；

……

活跃 15 天的用户，表示这个用户在过去 15 天中 15 天都启动了。

活跃天数越多的用户，其活跃程度越高，对 APP 的价值越大。我们可以用下面的图表作为呈现的方式（图 7-12），越往下，用户的活跃度越高。在数据统计中，一般我们认为活跃天数越多的用户，是最近常使用的用户，也是忠诚度较高的用户。发展稳定、留存良好的产品，随着用户忠诚度的提升和不断积累，活跃天数高的用户成分和占比也会逐步增加。

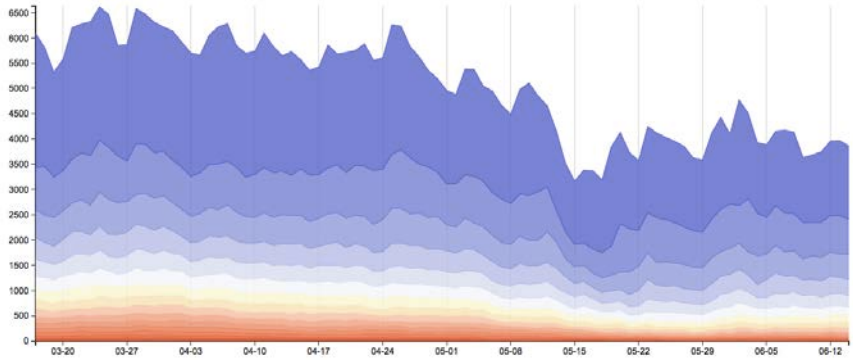



图 7-12

如果我们把用户的活动再进行细分，还会得到用户使用时长统计表和用户使用频率统计表。在用户使用时长统计表中我们会统计用户的单次使用时长、日使用时长。单次使用时长是指一次启动内应用在前台的持续时长（图 7-13）；日使用时长是指单设备在一天内使用时长的总和（图 7-14）。

单次使用时长分布明细

| 时长 | 启动次数 | 启动次数占比 |
|---|---|---|
| 1-3 秒 | 2419 | 13.2% |
| 4-10 秒 | 4245 | 23.2% |
| 11-30 秒 | 4430 | 24.2% |
| 31-60 秒 | 2538 | 13.8% |
| 1-3 分 | 3346 | 18.3% |
| 3-10 分 | 1305 | 7.1% |
| 10-30 分 | 43 | 0.2% |
| 30 分~ | 2 | 0% |

图 7-13

日使用时长分布明细

| 时长 | 用户数 | 用户数比例 |
|---|---|---|
| 1-3 秒 | 131 | 3.6% |
| 4-10 秒 | 239 | 6.6% |
| 11-30 秒 | 475 | 13.2% |
| 31-60 秒 | 415 | 11.5% |
| 1-3 分 | 814 | 22.6% |
| 3-10 分 | 1064 | 29.6% |
| 10-30 分 | 418 | 11.6% |
| 30 分~ | 38 | 1.1% |

图 7-14

使用频率一般会统计用户的日启动次数、周启动次数和月启动次数。日启动次数展示的是在查看日期内的用户每日启动次数的分布（图 7-15）；周启动次数展示的是查看日期前一个自然周内的用户每周启动次数的分布（图 7-16）；月启动次数展示的是查看日期前一个自然月内的用户每月启动次数的分布（图 7-17）。

日启动次数分布明细

| 启动次数 | 用户数 | 用户数比例 |
|---|---|---|
| 1-2 | 1716 | 44.4% |
| 3-5 | 974 | 25.2% |
| 6-9 | 547 | 14.2% |
| 10-19 | 440 | 11.4% |
| 20-49 | 172 | 4.5% |
| 50+ | 13 | 0.3% |

图 7-15

周启动次数分布明细

| 启动次数 | 用户数 | 用户数比例 |
|---|---|---|
| 1-2 | 4468 | 29% |
| 3-5 | 3928 | 25.5% |
| 6-9 | 2407 | 15.6% |
| 10-19 | 2707 | 17.6% |
| 20-49 | 1607 | 10.4% |
| 50-99 | 267 | 1.7% |
| 100+ | 40 | 0.3% |

图 7-16

月启动次数分布明细

| 启动次数 | 用户数 | 用户数比例 |
|---|---|---|
| 1-2 | 11955 | 20.1% |
| 3-5 | 14948 | 25.1% |
| 6-9 | 10489 | 17.6% |
| 10-19 | 11837 | 19.9% |
| 20-49 | 8202 | 13.8% |
| 50-99 | 1739 | 2.9% |
| 100-199 | 346 | 0.6% |
| 200-299 | 36 | 0.1% |
| 300+ | 14 | 0% |

图 7-17

在统计完用户活跃度之后面临的问题是用户能够活跃多久，这就是留存率的指标了。留存率指的是某段时间内的新增用户，经过一段时间后，仍继续使用应用的被认作是留存用户，这部分用户占当时新增用户的比例即为留存率。例如：5 月新增用户 200 人，这 200 人在 6 月启动过应用的有 100 人，7 月启动过应用的有 80 人，8 月启动过应用的有 50 人,则 5 月新增用户一个月后的留存率是 50%，二个月后的留存率是 40%，三个月后的留存率是 25%。用户留存率表如图 7-18 所示。

| 首次使用时间 | 新用户 | 1天后留存率 | 2天后留存率 | 3天后留存率 | 4天后留存率 | 5天后留存率 | 6天后留存率 | 7天后留存率 | 14天后留存率 |
|---|---|---|---|---|---|---|---|---|---|
| 2016/1/1 | 1423 | 18.5 | 7.7 | 5.3 | 3.6 | 3 | 3 | 2.8 | 1.2 |
| 2016/1/2 | 1315 | 23.2 | 12.2 | 7.2 | 6.1 | 5.3 | 3.3 | 2.2 | 1.3 |
| 2016/1/3 | 1808 | 39.9 | 18.4 | 12.4 | 11.1 | 7.9 | 6.3 | 4.9 | 4.1 |
| 2016/1/4 | 1648 | 36.3 | 18.4 | 15 | 9.6 | 7.6 | 6.5 | 5.1 | 2.9 |
| 2016/1/5 | 1381 | 34.6 | 20.8 | 12.5 | 8.8 | 6.6 | 6.1 | 4.1 | 2 |
| 2016/1/6 | 1267 | 39.4 | 18.6 | 11.2 | 7.3 | 7 | 4.7 | 4.5 | 1.4 |
| 2016/1/7 | 1571 | 37.1 | 16.5 | 11 | 7.5 | 5.9 | 4.6 | 4.5 | 2.2 |
| 2016/1/8 | 1515 | 36.9 | 17.4 | 9.2 | 8 | 5.7 | 6.2 | 4.6 | 2.7 |
| 2016/1/9 | 1504 | 35.6 | 17.4 | 10 | 8.2 | 6.4 | 4.3 | 4 | 2 |
| 2016/1/10 | 1483 | 35.7 | 16.5 | 9.1 | 8.4 | 7.3 | 6.3 | 6.3 | 2.4 |
| 2016/1/11 | 1460 | 34.7 | 17.3 | 11.6 | 7.9 | 7.8 | 5.2 | 4.9 | 1.6 |
| 2016/1/12 | 1317 | 35.8 | 18.6 | 12.1 | 9.1 | 7.1 | 5.2 | 4.2 | 2 |
| 2016/1/13 | 1305 | 36.3 | 16.2 | 10.9 | 6.7 | 5.6 | 4.4 | 4.2 | 1.9 |
| 2016/1/14 | 1590 | 35.5 | 16.1 | 10.7 | 7.1 | 5.7 | 4.5 | 4.2 | 2.1 |
| 2016/1/15 | 1669 | 35 | 15.3 | 10 | 6.8 | 5.5 | 5.9 | 4.6 | 1.9 |
| 2016/1/16 | 1225 | 38 | 19.8 | 11.3 | 9.3 | 7.1 | 7.3 | 5.1 | |
| 2016/1/17 | 1924 | 43.5 | 21.3 | 12.5 | 12.2 | 8.7 | 6.1 | 5.3 | |
| 2016/1/18 | 1540 | 38.2 | 20.9 | 15.8 | 10.7 | 8.2 | 5.8 | 5.5 | |
| 2016/1/19 | 1290 | 37.1 | 22.6 | 13.5 | 8.6 | 7.1 | 5.7 | 5 | |
| 2016/1/20 | 1328 | 37.7 | 18.8 | 11.8 | 7.8 | 6.3 | 5 | 5.6 | |

图 7-18

相较于留存率，用户新鲜度则是分析某日的活跃用户 DAU 的来源，将其分为当日新增用户、1 天前新安装并在当日启动的用户、

2 天前新安装并在当日启动的用户……30 天前新安装并在当日启动的用户、30+天前新安装并在当日启动的用户。这类图表我们一般使用颜色进行区分，可以简单地理解为颜色越浅的用户越新鲜，如图 7-19 所示。

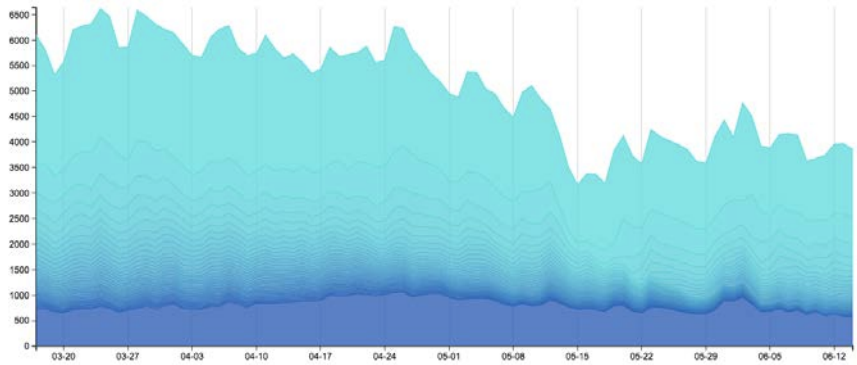



图 7-19

堆积图最上层为当日新增用户，依次往下分别为距离当日 1 天前新安装并在当日启动的用户、距离当日 2 天前……距离当日 30 天前、距离当日 30+天前新安装并在当日启动的用户。可以看到上层的用户是新用户，新用户的数量及占比跟应用推广行为相关。如果这个应用是新应用，或者最近推广力度非常强，新用户占比会比较大。下层的用户是老用户，老用户的数量及占比跟应用的质量和维持老用户的能力有关。如果这个应用处于稳定增长中，维系老用户的能力较强，老用户占比会比较大。用户新鲜度帮你从宏观上了解每日启动用户的新老用户比以及来源结构，观察留存的持续效果。从而制订正确的策略，在刺激新增与维系老用户之间建立平衡。

至此，日报、周报、月报这些数据统计的基本单元已经完成，同时对用户活跃度、留存率、新鲜度等数据的统计帮助我们更清晰地了解用户的活动表现。这些报表制作完成之后，以人工定期查询或是邮件自动触发等形式定期保质保量的发送，辅助监测业务各项指标的运作情况共同构成了我们的报表系统。

## 7.2 发现异常

报表系统搭建的目的在于对业务进行监控，所以对应的发现数据中的异常便是数据分析师需要掌握的第二项核心技能。不同的公司对数据时效性的要求不一样，但总体上以天为单位的数据日报是绝大多数公司的常态。

在上一节，我们帮助一家电商公司搭建了报表系统，下面就来看一看日常数据监控中会遇到的问题，在做异常数据监控分析之前，不妨先看看两个常用的概念同比与环比。

同比一般情况下指的是今年第 $n$ 月与去年第 $n$ 月比。同比发展速度主要是为了消除季节变动的影响，用以说明本期发展水平与去年同期发展水平对比而达到的相对发展速度。如，本期 2 月比去年 2 月，本期 6 月比去年 6 月等。其计算公式为：

$$\text{同比发展速度}=\text{本期发展水平}/\text{去年同期水平}\times 100\%$$

$$\text{同比增长速度}=(\text{本期发展水平}-\text{去年同期水平})/\text{去年同期水平}\times 100\%$$

在实际工作中，经常使用这个指标，如某年、某季、某月与上年同期对比计算的发展速度，就是同比发展速度。环比亦是属于统计术语，是本期统计数据与上期比较，例如 2015 年 7 月与 2015 年 6 月相比较，叫环比。

$$\text{环比发展速度}=\text{本期数}\div\text{上期数}\times 100\%$$

$$\text{环比增长率}=(\text{本期数}-\text{上期数})/\text{上期数}\times 100\%$$

环比增长率反映本期比上期增长了多少；环比发展速度，一般是指报告期水平与前一时期水平之比，表明现象逐期的发展速度。可以看到环比和同比是两个截然不同的概念，要注意区分。

图 7-20 是某公司近七日的 APP 注册用户数量，面对这样的数据图表，我们能看到哪些问题呢？

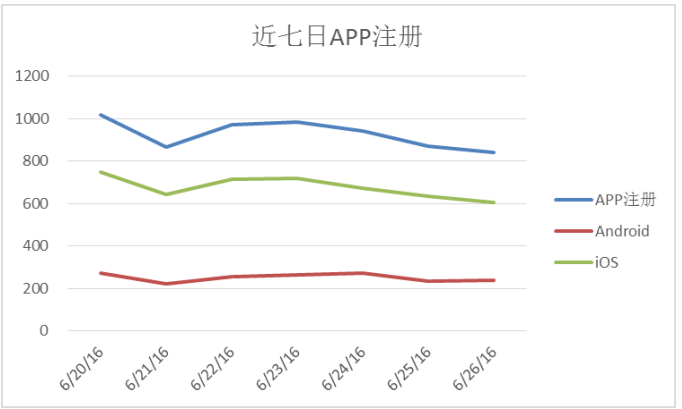



图 7-20

下面先简要概括下这张图表吧。iOS 和 Android 共同组成了 APP 注册的数据，总体上在过去的七天当中，6 月 21 日注册量同比下降较多，同时 6 月 23 日至 6 月 26 日，数据一直处于下降趋势。以上信息大概是我们能从这张图表中看出的主要信息了，那么这些信息中哪些是属于“异常数据分析”的目标呢？

如果我们直接看这张图，6 月 21 日显著降低是一个异常点；6 月 23 日至 6 月 26 日连续四天注册量下降又是另外一个异常点。其实这两个异常点都是基于之前说的一个概念：环比。将今天的与昨天的数据相比较，就是环比。如果我们再把上周的数据拿出来呢？将本周中每天的数据与上周中每天的数据进行一个同比，会有怎样的效果呢？

图 7-21 中我们把数据覆盖面提升到 14 天，是不是发现结论不一样了？之前的 6 月 21 日异常降低的数据点反倒与历史水平持平，大幅的下降是由于 6 月 20 日注册量异常增高造成的。同样，我们看到 6 月 15 日至 6 月 18 日这四天也是连续下降的，对照着日历我们很容易发现这几天对应着一周的下半周。站在这样一个角度我们看

到连续四天的下降是每周的周期性变动，不能算作异常点。目前的异常数据点是 6 月 20 日注册量的异常增高。异常增高主要是 iOS 下载注册异常增高带来的。这样问题就变成了：

为什么 6 月 20 日 iOS 的下载注册量异常增高？

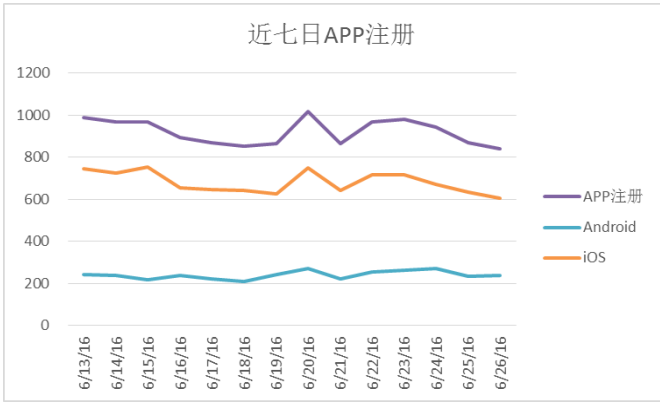



图 7-21

注册数据异常增高，我们有两个方向去追寻原因：一个是上游数据量变大了；一个是上游数据量没变但是转化率上升了。注册数据的上游数据是下载数据，我们看看这两周的下载数据波动情况如图 7-22 所示。

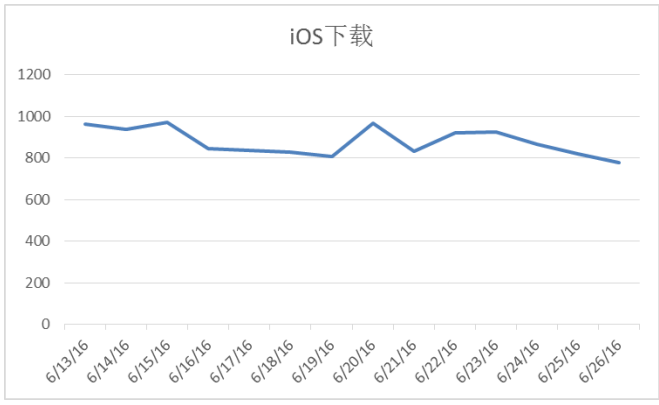



图 7-22

我们看到在 6 月 20 日 iOS 的下载量也出现了显著的增加，iOS 下载数据的上游是渠道。我们把渠道划分为有邀请码渠道和没邀请码渠道，邀请码渠道数据如图 7-23 所示。

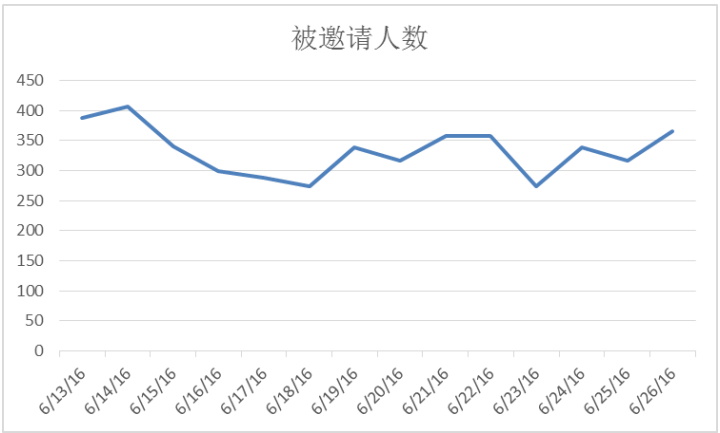



图 7-23

渠道邀请数据没有显著增加，那么增加的一定是非渠道邀请数据了。这时我们可以问问市场部是否在非邀请码渠道做过投放？市场部反馈在 6 月 20 日做过应用市场搜索优化（App Store Optimization，ASO），至此原因就找出来了。另一方面也验证了 ASO 优化的效果是显著有效的。

这个异常数据分析的案例是给大家提供一种解决问题的思路，事实上，当市场部进行 ASO 优化时应当首先通知数据部门，这样当数据分析师看到这个波动时心中便知道原因是什么了。相反，如果数据没有波动反而可能是异常的。毕竟我们做了前期投入但是没有变化这也算是不合理的数据表现了。

我们再看看下面的数据，这是某电商交易公司的日交易单量（图 7-24）。

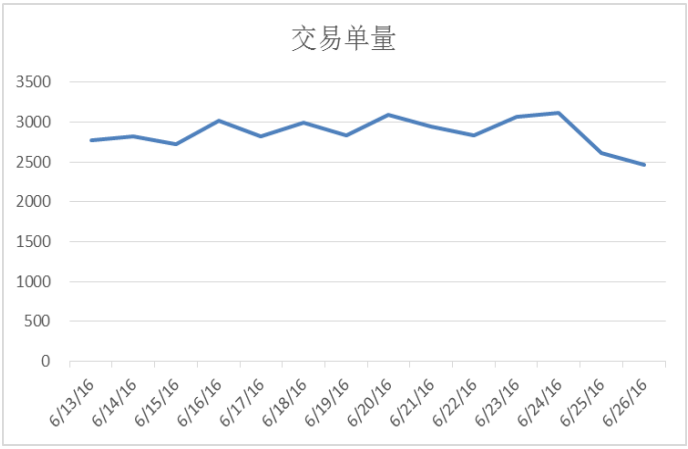



图 7-24

大家能够一眼看到的问题是 6 月 25 日和 6 月 26 日的交易量出现了大幅的下降，除了这个问题大家还能发现别的问题吗？

不知道大家是否关注“6.18”时各家电商网站做的各类推广活动，“6.18”作为继“双十一”之后又一个人造节日，许多电商广告会在这一天集中发力，用户也会在这一天产生各种非理性消费订单，作为一家电商公司在“6.18”这一天理论上交易量是应该上涨的，不上涨是不是一个异常点呢？这就是异常数据分析面临的第二个重要而又核心的问题：

**应该异常的数据不异常！**

针对上面的问题我们回顾该电商公司在“6.18”的运营活动，发现该公司在“6.18”做了产品活动宣传，但是为什么没有实质性的数据增长呢？我们去追寻交易的上游数据，也即：有效订单和临时订单（图 7-25 和图 7-26）

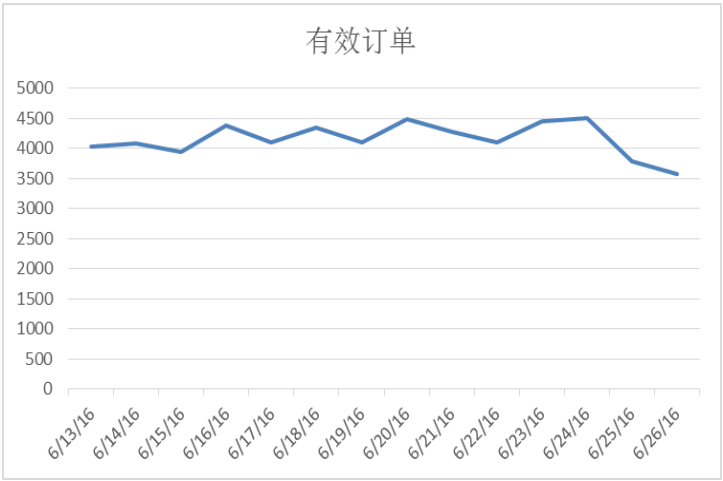



图 7-25

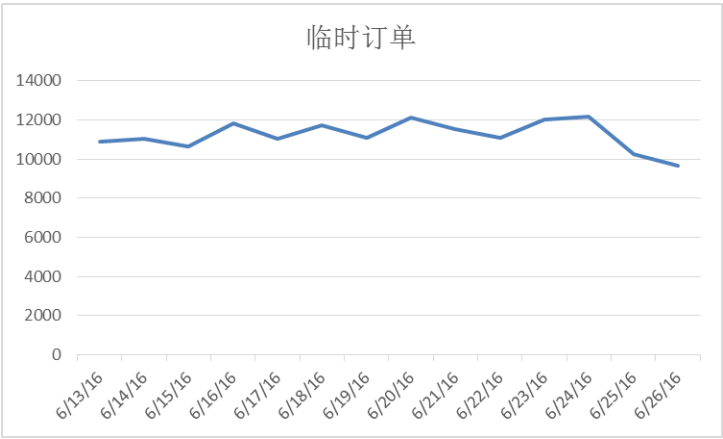



图 7-26

我们发现临时订单和有效订单的整体趋势和成交订单大体相同，那么就排除转化率的问题。在临时订单的上游是什么数据呢？是订单来源的渠道！我们去查询下交易来源的几个渠道的数据表现，如图 7-27 所示。

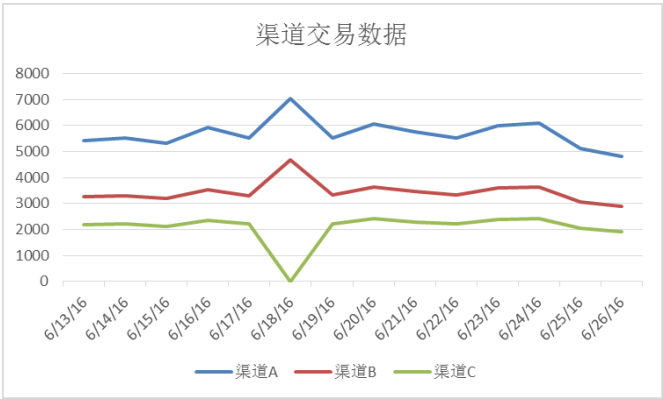



图 7-27

问题一下就清楚了，渠道 A 和渠道 B 的交易量都因为“6.18”而出现较大幅度的上升，但是渠道 C 在“6.18”这天的数据量突然变成 0。我们要求商务部对渠道 C 进行问责，为什么在“6.18”当天渠道出现问题？渠道 C 表示当天量太大导致了部分服务器宕机了，愿意在接下来的合作中降低价格作为损失补偿……

这个异常数据分析到这里算是完成了绝大部分，如果我们能够再总结一下：这个问题的来源是渠道宕机，渠道宕机我们没法预料，但是接下来可以有预警机制，诸如在重大节日之前和各个渠道提前通知做好准备，实时数据监控以及出现问题后的备选方案是什么？这些算是我们异常数据分析最后锦上添花的事情。这些充满附加值的任务完成好坏往往是数据分析师更进一步的关键与核心，每件事情都不一样，大家尽情地发挥自己的主观能动性吧！

## 7.3 数据需求

数据需求是数据分析师日常工作中的一个常态性、插入性任务。说常态性是因为数据分析师几乎每天都要处理各种数据需求，说插入性任务是因为需求常常是急迫性的、打断性的。数据需求常常需要你要在多长时间内做出来，用于做某项决策。

小 A 上班后收到的第一份数据需求邮件是这样的（图 7-28）。

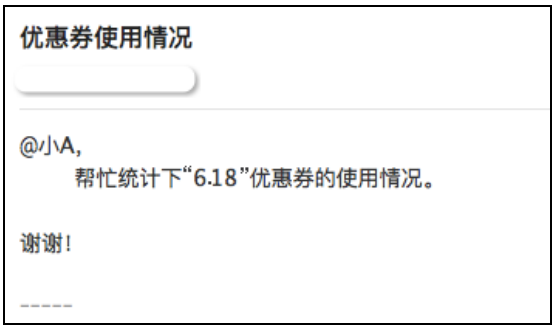



图 7-28

面对这样的需求你是不是一头雾水？统计使用情况应该统计哪些内容？细细一想，应该是：

（1）发放了多少张优惠券

（2）使用了多少张优惠券

（3）使用占比

于是小 A 回复了邮件（以下数值仅供参考，不代表任何公司的任何业务），如图 7-29 所示。


图 7-29

然后问题又来了（图 7-30）。

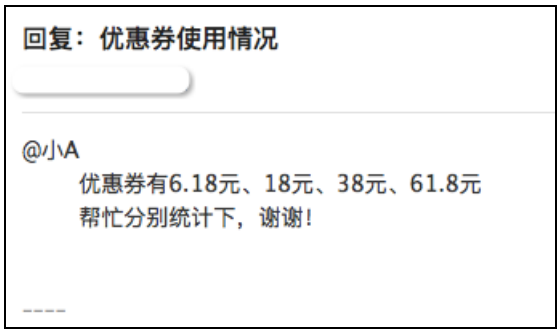



图 7-30

这样一边提问题一边回复邮件的方式处理数据需求是许多新手都会遇到的问题，别人提问题我们解决问题，没什么错。但是损失的是你自己的时间，实际的问题能否被解决还要另当别论。

当然，还有另一类比较专业的数据需求（图 7-31）。

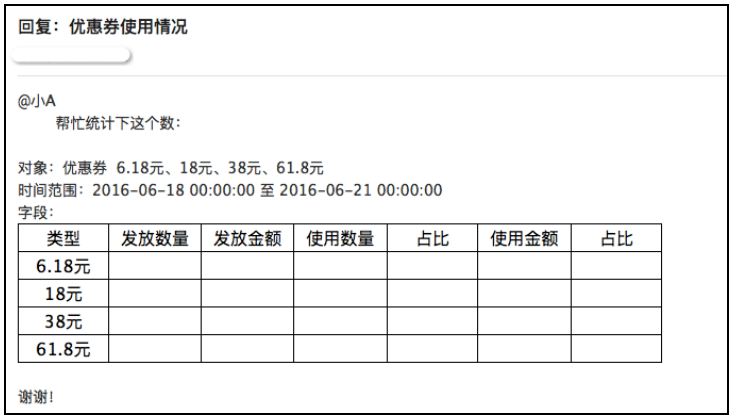



图 7-31

这种数据需求就相当专业了，思路清晰、结构明了，数据分析师只需要填好表格就好了。

一个是有来有回的讨论需求，一个是命令直达的完成需求，显然第二种方式更加高效直接，但是作为数据分析师的你知不知道你提的需求是否解决了问题呢？你不知道，所以你只是个执行者，服

从命令的执行者。作为执行者的一个好处就是磨炼技能，能在短时间内提升 MySQL 查询语句的水平，但是随着查询语句掌握得越来越熟练，你会发现自己进入了一个瓶颈期。数据需求这么大量的插入性工作难道只是用来练练手？简单的数据需求如何做得更好，更好地帮助别人解决问题？这个时候你就需要知道数据需求方的目标了！

**这个数据需求的目的是什么？**

这个问题要贯彻在数据需求的始终，如果知道对方的目的是什么，你就能够不仅把思维停留在 MySQL 查询语句怎么写的问题上，你会开始关注如何帮助别人解决问题，这才是核心中的核心。

回到刚刚的优惠券数据需求，我们问一问数据需求方：你的目的是什么？

需求方：我们想要评估一下上次发放优惠券的效果，若效果不错，那么 7 月会再做一次优惠券发放活动。

这样我们再来看一看这个需求，他的核心目的是评估之前活动的效果，为了下一次的活动做铺垫。那么我们需要做的就变成了评估上次活动的效果了，进一步来说就是活动的 ROI，也即投入产出比。那么从 ROI 的角度，我们需要统计哪些字段呢？

首先是成本，我们花了多少钱？

排除掉人力成本，我们的成本核心在于使用优惠券的用户，他们减免的金额是我们的活动成本，这时我们需要统计实际使用的类型与张数。为了给予大家一个直观的印象，我们不妨从 MySQL 查询语句开始做这个需求：

```
    select coupon_type 优惠券类型,count(coupon_id) 优惠券张数,sum(coupon_amt) 优惠券金额
    from user_coupon
    where coupon_status='success'and create_time between '2016-06-18 00:00:00' and '2016-06-21 00:00:00'
    group by coupon_type
```

通过之前对 MySQL 的学习，上面的语句应该很容易理解，我们得到如表 7-2 所示的结果。

表 7-2

| 优惠券类型 | 优惠券张数 | 优惠券金额 |
| --- | --- | --- |
| 6.18 元 | 673 | 4159 |
| 18 元 | 329 | 5922 |
| 38 元 | 172 | 6536 |
| 61.8 元 | 61 | 3770 |

这样我们得到每个优惠券的成本情况，总成本统计可以在 MySQL 中加一个嵌套：

```
select sum(coupon_amt) 成本
from(
select coupon_type 优惠券类型,count(coupon_id) 优惠券张数,sum(coupon_amt) 优惠券金额
from user_coupon
where coupon_status='success'and create_time between '2016-06-18 00:00:00' and '2016-06-21 00:00:00'
group by coupon_type
) a
```

上一段的 MySQL 查询语句使用了一个嵌套，把之前的查询语句（标灰）得到的结果作为一张表命名为表 a，我们对这张表再次进行统计和查询得到如下结果。

| 成本 | 20387 |
| --- | --- |

粗略统计我们此次活动的总成本是 20387 元。

成本统计之后我们再看看本次活动的效果，效果体现的核心在于利润的提高，如果这些优惠券发放之后带来了交易量的提升，从而也就带来了利润的提升，这就是最直观的活动效果。与此同时，考虑到用户的持续消费能力，新增加的用户会在后续为我们带来收益，因此新增用户也加入本次活动效果的统计。通过 MySQL 查询

数据库中的新增用户和提升的交易量。

```
select date(create_time) 日期,count(user_id) 注册人数
from user
where create_time like '2016-06%'
group by 日期
```

上一段 MySQL 查询语句使用了一个新的函数“like”和通配符“%”，一般情况下，它们都是一起出现，表示“create_time”前面 7 个字符匹配“2016-06”的任意日期都会满足条件，即 2016 年 6 月整月都会被调取，得到如表 7-3 所示结果。

表 7-3

| 日　　期 | 注册人数 |
|---|---|
| 6 月 1 日 | 5203 |
| 6 月 2 日 | 5105 |
| 6 月 3 日 | 5068 |
| 6 月 4 日 | 5213 |
| 6 月 5 日 | 5153 |
| 6 月 6 日 | 5306 |
| 6 月 7 日 | 5343 |
| 6 月 8 日 | 5082 |
| 6 月 9 日 | 5321 |
| 6 月 10 日 | 5138 |
| 6 月 11 日 | 5041 |
| 6 月 12 日 | 5311 |
| 6 月 13 日 | 5390 |
| 6 月 14 日 | 5374 |
| 6 月 15 日 | 5306 |
| 6 月 16 日 | 5303 |

续表

| 日　　期 | 注册人数 |
| --- | --- |
| 6 月 17 日 | 5131 |
| 6 月 18 日 | 5778 |
| 6 月 19 日 | 5235 |
| 6 月 20 日 | 5332 |
| 6 月 21 日 | 5296 |
| 6 月 22 日 | 5129 |
| 6 月 23 日 | 5192 |
| 6 月 24 日 | 5324 |
| 6 月 25 日 | 5294 |
| 6 月 26 日 | 5217 |
| 6 月 27 日 | 5194 |
| 6 月 28 日 | 5149 |
| 6 月 29 日 | 5286 |
| 6 月 30 日 | 5114 |

然后再统计交易量的变化情况。

```
    select date(a.create_time) 日期,sum(a.ord_amt) 成交金额,sum(b.price) 原价, sum(a.ord_amt)-sum(b.price) 利润
    from trx_order a
    left join product b on a.product_id=b.product _id
    group by  日期
```

上一段 MySQL 查询语句在查询字段中增加了计算“sum(a.ord_amt)-sum(b.price)”，前者是订单价格，后者是商品原价，同时我们还用了“left join”将“product”这张记录产品原始价格的表通过“product *id*”与“*trx*order”连接起来形成一一匹配，这样每一条交易记录都有该产品的原价了。得到如表 7-4 所示结果。

表 7-4

| 日　　期 | 交易额 | 成　　本 | 利　　润 |
|---|---|---|---|
| 6 月 1 日 | 336757 | 306449 | 30308 |
| 6 月 2 日 | 324176 | 295000 | 29176 |
| 6 月 3 日 | 332697 | 302754 | 29943 |
| 6 月 4 日 | 336722 | 306417 | 30305 |
| 6 月 5 日 | 336889 | 306569 | 30320 |
| 6 月 6 日 | 339162 | 308637 | 30525 |
| 6 月 7 日 | 336517 | 306230 | 30287 |
| 6 月 8 日 | 327799 | 298297 | 29502 |
| 6 月 9 日 | 323491 | 294377 | 29114 |
| 6 月 10 日 | 330856 | 301079 | 29777 |
| 6 月 11 日 | 327037 | 297604 | 29433 |
| 6 月 12 日 | 325010 | 295759 | 29251 |
| 6 月 13 日 | 336514 | 306228 | 30286 |
| 6 月 14 日 | 337418 | 307050 | 30368 |
| 6 月 15 日 | 331569 | 301728 | 29841 |
| 6 月 16 日 | 338936 | 308432 | 30504 |
| 6 月 17 日 | 336567 | 306276 | 30291 |
| 6 月 18 日 | 517446 | 470876 | 46570 |
| 6 月 19 日 | 466220 | 424260 | 41960 |
| 6 月 20 日 | 397749 | 361952 | 35797 |
| 6 月 21 日 | 332074 | 302187 | 29887 |
| 6 月 22 日 | 322412 | 293395 | 29017 |
| 6 月 23 日 | 332064 | 302178 | 29886 |
| 6 月 24 日 | 335983 | 305745 | 30238 |

续表

| 日　　期 | 交易额 | 成　　本 | 利　　润 |
|---|---|---|---|
| 6 月 25 日 | 327736 | 298240 | 29496 |
| 6 月 26 日 | 326397 | 297021 | 29376 |
| 6 月 27 日 | 323737 | 294601 | 29136 |
| 6 月 28 日 | 322795 | 293743 | 29052 |
| 6 月 29 日 | 338743 | 308256 | 30487 |
| 6 月 30 日 | 338949 | 308444 | 30505 |

大家不妨发挥下自己的数据敏感度，“6.18”当天我们的活动收益是多少呢？

如果直接看数字不够直观，那么可以画个折线图，如图 7-32 和图 7-33 所示。


图 7-32

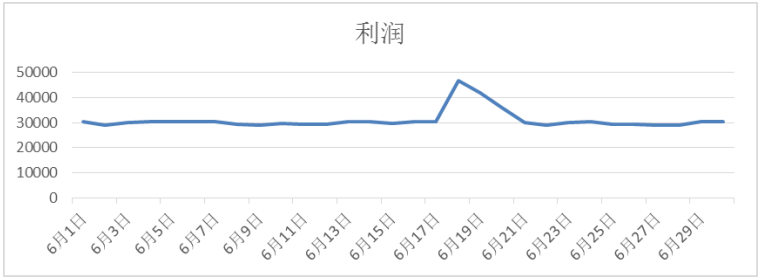



图 7-33

我们很容易在 6 月 18 日附近看到一个波峰，严格意义上来说，我们可以做个 T-Test 来检验这个波峰的显著性，但是明显这个波峰的高度可以百分百确定它的显著性。接下来需要把这个波峰量化出来，计算方法是：

差值=波峰值－非波峰均值

这种方法比较简单粗暴，但是切实有效。这样就计算得出 6 月 18 日当天的活动为我们多带来了 553 个新增用户和 34738 元的附加收益。按照历史上平均获客成本为 5 元/人，本次活动的收益为：

553×5+34738=37503 元

相较于我们的成本 20387 元，本次活动的收益比为：

(37503–20387)/20387=84%

总的来看，上次活动的效果相当好，收益率高达 84%，看来这样的活动可以多做几期（以上数值均仅供参考，不代表任何公司的任何业务）。

至此本次数据需求就完成了。是不是比想象中要复杂很多呢？

数据需求的核心一定是帮助别人解决问题实现目标，所以做需求的时候一定要问对方你的目的是什么？了解目的之后分析人员要能够根据对方的目的来了解他的需求，然后再到数据库中寻找数据支撑，最终实现数据需求。

## 7.4 项目分析

在之前的章节我们把项目分析分成主动项目分析和被动项目分析，在实际工作的案例中，项目分析又可以区分为策略分析和项目推进。策略分析倾向于给出方案和指导建议，通过邮件或者其他方式给出分析报告和指导建议。项目推进需要我们用数据知道决策的

同时把项目完成和推进。前者是谋划而不执行，后者不仅需要谋划还需要执行配合。

首先我们来看一个策略分析的分析报告。

【分析目的】

提升理财产品销量。

【探究问题】

（1）购买理财产品用户的属性分布；

（2）购买理财产品用户的行为；

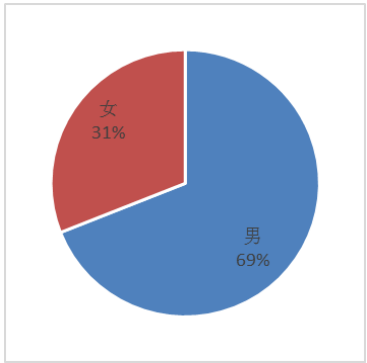
（3）未购买理财产品用户的属性；

（4）未购买理财产品用户的行为。

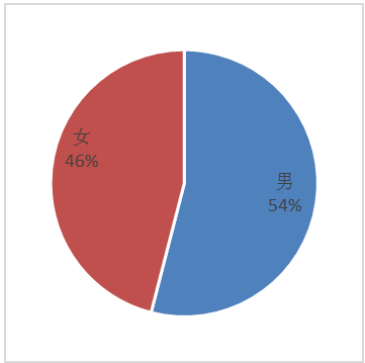
【分析过程】

购买理财产品的用户属性分布：用户静态的不以时间为转移的数据，关注用户的性别、年龄、所在地域以及设备类型。

- 性别比例中购买理财产品的用户以男性为主，男性用户是女性用户的两倍，而未购买理财产品的用户中女性比例提升，说明女性用户的购买积极性略低于男性用户的购买积极性（图 7-34）。
- 城市分布中用户主要分布在二、三线城市，同时四线城市在未购买用户中的占比相对较高，说明四线城市的用户活跃度显著低于二、三线城市的用户活跃度（图 7-35）。
- 年龄分布中用户主要集中在 23 岁~27 岁的群体，购买用户与未购买用户的年龄分布基本相同（图 7-36）。

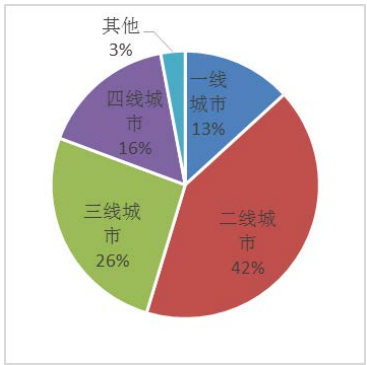



购买用户

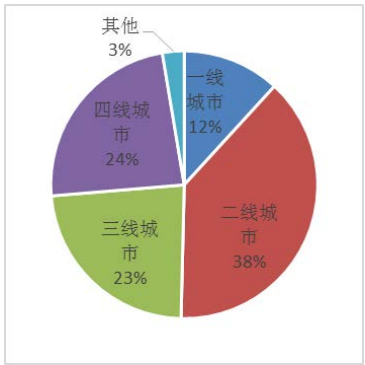



未购买用户

图 7-34

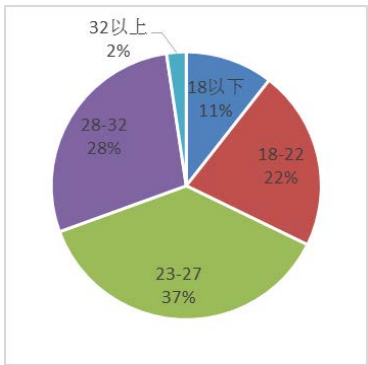



购买用户

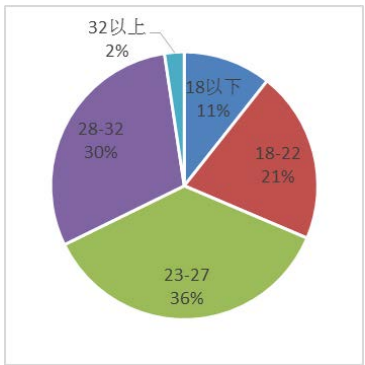



未购买用户

图 7-35




购买用户




未购买用户

图 7-36

- 用户设备分布上用户主要使用 iOS 设备，同时 Android 用户在未购买用户中的占比较高，说明 Android 手机用户的购买活跃度低于 iOS 手机用户的购买活跃度（图 7-37）。

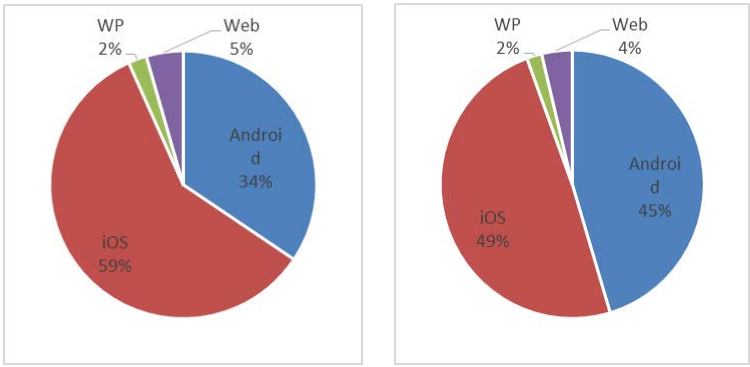
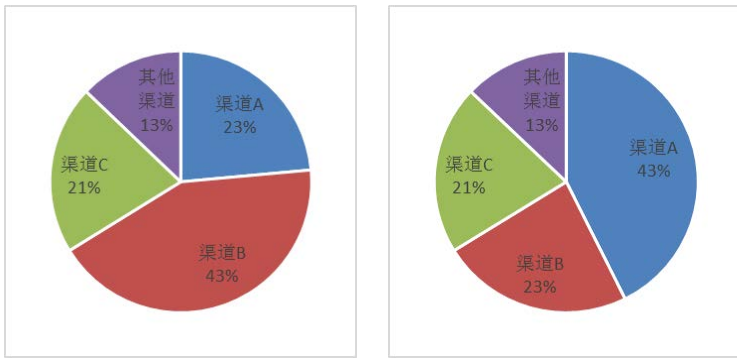



图 7-37

综上，我们的理财产品用户主要是在二、三线城市使用苹果手机的 23 岁~27 岁的男性群体。

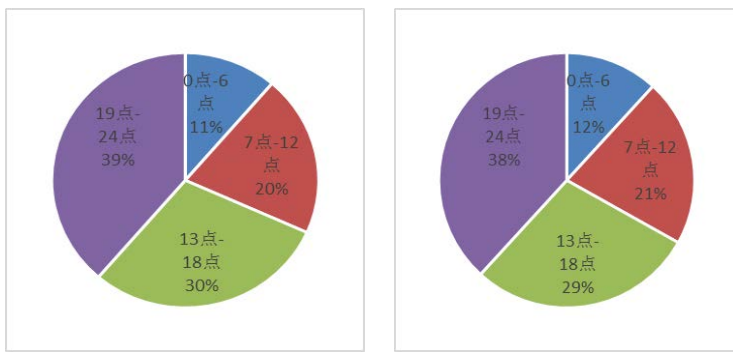
购买理财产品的用户行为数据：用户在本平台留下的行为类数据，反映用户的操作意图和行为特征。

- 我们的购买用户主要分布在渠道 B，未购买用户主要分布在渠道 A，渠道 B 的用户显著比其他渠道的用户要活跃（图 7-38）。
- 用户的注册时间主要集中在 19 点~24 点，购买用户与未购买用户的分布大体相同，说明用户主要在晚间进行注册活动（图 7-39）。
- 购买用户的主要绑卡类型是信用卡，未购买用户的主要绑卡类型是储蓄卡，说明信用卡用户的活跃度显著高于储蓄卡用户（图 7-40）。



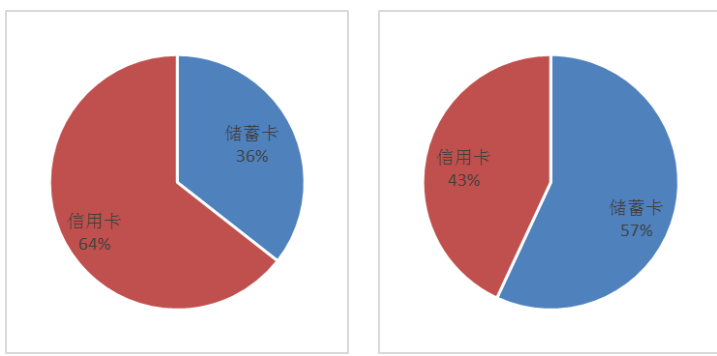

购买用户　　未购买用户

图 7-38




购买用户　　未购买用户

图 7-39




购买用户　　未购买用户

图 7-40

【数据结论】

- 购买理财产品的用户以男性为主（69%），女性用户积极性不高。
- 购买理财产品的用户主要分布在二、三线城市，一线及四线城市购买量一般。
- 购买理财产品的用户多使用苹果手机。
- 购买理财产品的用户主要来源于渠道B,渠道A的用户质量较差。
- 用户一般在晚上 6 点~12 点进行注册。
- 绑定信用卡进行理财产品购买的用户显著高于绑定储蓄卡的用户。

【可行性方案】

（1）在用户推广时主要面向二、三线城市使用苹果手机的男性群体。

（2）增加渠道 B 的投放，对渠道 A 进行优化或是成本控制。

（3）营销活动可以将时间定在晚上 6 点~12 点，此段时间用户较为活跃。

（4）增加对信用卡用户的引导和激励。

至此，一份简单的分析报告就结束了。梳理这份报告，我们看到一份基本的分析报告至少包含的结构如图 7-41 所示。

这是一个层层递进的关系，从确定目标开始把这个目标拆分成几个需要分析和讨论的问题，然后对这些问题分别进行探讨和分析就是分析过程，分析一定要有结论和总结，最后根据这些结论和总结形成执行方案。与此同时，在制作分析报告时常常需要做对照组，一个是我们已经有的好的样本：已经购买理财产品的用户；一个是我们需要优化和改进的样本：还未购买理财产品的用户。我们比较这两个样本在用户属性和用户行为两个维度之间的差异，找出比较明显的结论，最后针对这些结论制订可以优化样本的执行方案，完成分析报告。

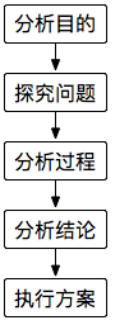



图 7-41

接下来，我们再来看一个有趣的项目推进案例。假设你接收到一个指标：一个月内卖出 1000 份保险。别管丧心病狂的老板是如何制订这个目标的，假设他是合理的，你该怎么办?

事实上，在接到这个问题的时候一般人是完全没有思路的，完全没有头绪的情况下如何下手？当你冷静下来之后你发现其实这是一个营销的问题。营销我不太懂啊怎么办？于是你抬头看了看你的书架，你发现了一本《营销宝典——三分钟搞定客户》。假设它是存在的，于是你尝试着去了解一下如何做营销。

这时候你看到了各种成功案例：比尔·盖茨如何成功？你该如何学习他！安妮·马尔卡希如何成功？你该如何学习他！汤姆·霍普金斯如何成功？你该如何学习他……这些还不够！这些人太厉害了，没法直接学。于是延伸到从他们案例里抽离出来的方法论：客户心理学、5W2H、马斯洛人性管理、演讲与口才……实践是检验真理的唯一标准，你抱着一堆传单和材料上街了。一方面你在想：没用的!你不可能在一个月内卖掉 1000 份保险;另一方面你又在想：其实不一定，如果有钻研精神说不定一个月就卖掉了……

以上只是一个传统营销手段的抽象概括，目前市场上卖保险的多是采用这种成功学引进门，各种营销方法与课程的学习然后给你一大堆传单开始上街发传单去，当然这些保险从业人员的第一笔订单多来源于亲戚朋友，能否取得效果就见仁见智了。

下面，我们不妨换一种思路和方法来卖保险，来谈谈如何把卖保险这件事情科学化与数据化。

你已经知道要卖的是什么，那么接下来不妨思考第一个问题：在哪里卖？微信朋友圈、公众账号、人人网、QQ 空间、漂流瓶（这个渠道真有打广告的）、淘宝网、自己开家网店（呵呵）、人民广场发传单……身边的渠道很多，请理清楚它们，然后用一天的时间在网络渠道进行资源投放，等待三天，这三天里进行线下推广（发传单）。线下推广请详细记录推广反馈情况，诸如：接受你传单的人群年龄分布（目测）、穿着打扮、结伴情况……同时，这三天时间请采用不同的营销方式，第一天职业化西装领带装扮，第二天休闲化运动休闲装，第三天精英化弄一身“老板”装……

三天之后请统计所有渠道的点击量（咨询量），以及转化情况，然后选出优质渠道。这里你可能遇到一个问题，那就是一个人的精力太有限，忙不完这些渠道，你可以找人啊！完成 1000 份保险的收益不少，可以使用股权激励机制，于是有了电子商务。

假设你已经筛选出了几个优质渠道（渠道总量占比超过整体的 80%），这时候你需要维护这些渠道。假设人民广场卖保险这个渠道也是优质的，这个渠道也得优化。在正式开始之前请先明确目标，假设 30 天时间，我们已经浪费了 4 天，还剩下 24 天。姑且算之前的三天测试期卖了 50 份保险，还剩下 950 份保险要在 24 天完成，任重而道远，但我们没时间了。

互联网渠道的页面层级转化率之类都比较好统计，这里不做赘述，假设互联网能完成目标的××%。下面来谈谈人民广场发传单的事情，假设你穿“老板”装的推销量最好，职业装其次，休闲装最差，那么请计算下投入产出比，能不能永久搞到一身“老板”装，如果不行能不能借？如果不能天天借那么请在每周的人流高峰期时间借（何时为人流高峰期自己想）。现在好了，你明确了周一到周五穿职业装，周末穿“老板”装，假设你的目标客户是 50 岁左右、多人出行、用苹果手机拍照、戴墨镜、大叔……我们假设接下你的传

单就产生一个临时单，咨询你问题产生一个有效单，填写个人信息产生一个正式单，付钱买保险产生一个付款单，好了，你可以计算转化率了。想想你的人民广场目标 950×（1-××%），看看你每天的单量、转化率及增长率，你该如何实现目标？

方案一：增加临时单量

方案二：提高转化率

为了增加临时单，你开始 6 点起床，凌晨 2 点回家，为了提高转化率，你啃着面包反思总结今天的收获和明天的方案。可是你觉得临时单还是太少，又用股权激励机制招了几个销售员并教会他们如何使用“老板”装秒杀客户。你觉得应该找一个电话回访人员来提高转化率，于是你用股权激励机制招了一个客服专员。

接下来的几天，收益速度有明显提高，但是还是不够，离 1000 份的目标还是有差距，但是我们不能放弃，怎么办？送！免费送！用你未来得到的利润往里面贴，抽奖送，送多少，你可以说送 1000 套，具体送多少套只有你自己知道！宣传效果还不好？“造节”！×月×日全球保险日，送××保险公司制作的安全套！你不觉得×××小了点吗？这是要成立一个安全套公司吗？不是！这是战略！可还不够？怎么办？挂个喇叭：“老板跟小姨子跑了，原价 8999 元的保险现价 899 元!”恶趣味炒作。再不行怎么办？雇佣一个城管来砸你的保险销售人员，义愤填膺的群众会分分钟把你推向高潮。

好了，你已经成立了一个创业者团队了。有人在搞电子商务，有人在搞客户服务，有人在搞线下推广，有人负责炒作宣传……这时，你可以告诉他们：完成目标，我给你们总利润的 60%。接下来你可以用你的股权激励招募更多的销售人员，给他们配备“老板”装。再找一个产品经理，让他帮你出主意提高转化率。

在这个时候，你发现自己好像成立了一家公司。

回顾整个过程，其实一个月卖出 1000 份保险的本身就是一个项目，然后为了完成这个项目你不断地获取资源来实现目标，在这样

的一个过程中，你拓展渠道、监控渠道、优化渠道，提升转化率、成交率，控制收益与成本……全都是数据在支撑你推进项目，这时候与其说在做项目性分析，不如说你是项目经理。所以数据分析的进阶形态就是项目经理，完成一个项目所产生的收益是实实在在为公司带来利润，这时衡量你自己的标准不再是上班打卡考勤执行力，而是你能够为公司带来多大的收益。

你是否有过这样的经历：辛辛苦苦加班加点做出了一个方案，满怀信心的提交给你的领导，他只是匆匆看了几眼然后说一句：我觉得这里不科学！我觉得那里有欠缺！我觉得客户不会喜欢！我觉得……

太多的“我觉得”，太多的“应该是”，太多的“假设顾客是”，是否让人无从辩解，任何东西都有正反两面，从不同的角度会看出不同的结果。喜欢改变和创新的人看到它的优点，保守求稳的人看到它的缺点，我们会说这个方案的优势大于劣势让我试一试，而领导看到了它的风险和损失，不愿让你尝试，那么我们该如何让方案更加具有说服力呢？

我会说：用数据！让我们的双脚更加坚实的站在大地上。

大数据时代的今天，数据信息铺天盖地，该如何用数据为自己服务呢？不同的行业有不同的用法，这里就用一个简单地比喻来讲述数据的思想。

**问：如何设计一个凳子？**

方案一：照着别的凳子的模样复制一个。

方案二：找来所有使用起来十分舒服的凳子，量取各项数据，取平均值，做一个凳子。

方案三：从人体构造角度，重新设计一个凳子。

采取方案一的人注定只能当个工人；

采取方案二的人会是一个小老板；

采取方案三的人创造了“苹果”。

这只是个小插曲，乔布斯具体怎么做的我也不清楚，那么我们就以最普通的角度阐述一下用数据如何实现方案三。让我们一起设计这个小凳子。

不如从这个凳子的长宽高开始（我们不妨认定这个凳子的表面是长方形）。

这个凳子该多长？这是一个问题。如何解决？看其他凳子的长度吗？那不如采用方案二。

既然我们要设计一个完美的凳子，怎么能照搬照抄呢？因此我问我自己，凳子的长度会影响用户的什么体验？

当然是屁股。那么凳子长度的数据来源就有了，屁股的大小。可是屁股的大小不好判定也不好取样测量，如何解决这个问题？能不能找到其他能客观描述的数据？

能！这里我第一个想到了髋骨的大小！这个数据在生物学领域是可以找到的。我们姑且不论这个数据找到的难度大小，谁让我们要做最完美的凳子呢。

好了，我们找到了一大堆髋骨宽度的数据，最直观的统计当然是取平均值。我们会对目标人群进行分类，因为男人的屁股、女人的屁股和小孩的屁股差别总是有的。但是板凳的宽度取平均值会带来什么后果？近百分之五十的人屁股会比凳子大，这是一个失败的设计。因此我们要加长，加多少？这个数据怎么得来？不能凭空捏造或者凭感觉加个 30 厘米，这里我想到了六西格玛策略（来源于正态分布，不懂的自行百度）。

求出原有数据的标准差，在均值上增加三倍的标准差即可满足99.7%的人的屁股大小。是不是很合理？能不能让人信服？能让人信服但不是很合理。我们必须考虑板凳两边留余的问题，不然别人稍微偏一点点就会很不舒服。但是我们回过头来看刚刚的均值加三西

格玛的板凳长度，一定会比很多人的屁股都大。这样一来我们是不是要考虑缩小成二西格玛？不能这么草率，我们要制造最完美的凳子！这里我采用的方法是抽样调查，就是要随机抽样目标人群来体验这个板凳的长度，现在目标已经定在了均值加三西格玛附近，找一个合适的长度会不会简单很多？

我们用如此科学的方法定下了板凳的长度，谁还会质疑说这个板凳是不是太短了？如果客户的屁股比较大怎么办？你如何回答？看看之前的数据：99.7%！你还没信心说服他吗？

花了这么大的功夫我们总算把板凳最基本的框架里的其中一条初步定下来了，之后若有更科学的方法解决三西格玛的问题我们继续讨论！这条凳子还剩下的宽、高、板凳几个腿、材料、平凹，等等，我们不急，慢慢来。

项目分析的环节就到此结束了，这是一个数据分析师的看家本领和核心技能，不知道大家对项目性分析有什么看法呢？

【拓展阅读】

项目分析的内容多种多样，在项目分析中，有时候会需要一些建模模型来帮助我们进行决策。下面是基于逻辑回归使用 R 语言预测用户是否会购买某类产品的一个模型搭建的简化过程。

假设我们有一批用户的数据，对于这些用户我们已经知道他们是否购买了产品。我们想要通过这些数据来建立一个模型，这样可以预测新来的用户会不会购买产品，使用 Logistic 回归建模的方法来进行这个模型搭建。

**第一步：数据调取与数据清洗**

（1）使用 MySQL 将用户属性数据和用户行为数据分别调取

（2）将性别、城市、学校等字符型数据分级量化

（3）将用户行为数据分箱

第一步往往是最耗费时间的一个环节，这个环节的连续型变量的分箱处理和字符型变量的量化标准需要耗费许多时间。

**第二步：数据准备**

假设我们通过第一步的数据清洗得到了两个数据集，接下来我们要对两个数据集进行合并和抽样，为我们的模型搭建做准备。

```
iddata=read.csv('sqlresult_id.csv',header=T)#用户属性数据
urdata=read.csv('sqlresult_ur.csv',header=T)#用户行为数据
alldata=merge(iddata,urdata,by='user_id')#数据处理
rdn=1500#样本量控制
rd=sample(1:nrow(data8),rdn,replace=FALSE) #随机选择
traindata = alldata[rd,] #训练数据
testdata = alldata[-rd,] #测试数据
```

**第三步：回归建模**

通过第二步的数据准备我们已经有了建模的原始数据，接下来用 R 语言的 glm 逻辑回归函数对数据进行建模。

（1）选取变量进行全变量建模

```
fit.full=glm(due~.,data=datatrain,family=binomial())#建模函数
summary(fit.full)#查看建模结果
```

（2）寻找显著性较强的变量

建模的结果中需要重点关注这样几个指标：

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’0.05 ‘.’ 0.1 ‘ ’ 1

其中“0”、“0.001”、“0.01”、“0.05”、“0.1”、“1”代表变量的 P 值

我们将标注“.”、“*”、“**”、“***”的变量再次选入模型进行建模。

（3）显著性变量重新建模

```
    fit.reduced=lm(due~v1+v2+v3+v4+v5+v6+v7+v8,data=data
train,family=binomial())
    summary(fit.reduced)
```

我们可能会得到如下结果：

```
    Coefficients:
              Estimate Std. Error t value Pr(>|t|)
    (Intercept)  0.829679  0.050341  16.481  < 2e-16***
    V1           0.013531  0.006270   2.158  0.0319 *
    V2          -0.019255  0.004289  -4.489 1.10e-05***
    V4          -0.008961  0.002258  -3.969 9.47e-05***
```

这个结果可能通过以上两个建模环节就能得到，而更多的时候要反复地进行变量选入与剔除，直到选择的变量都显著为止。

**第四步：模型检验**

接下来介绍几个用来衡量模型好坏的指标，它们分别从不同的维度去衡量了一个模型的好坏优劣，如果指标达不到要求则需要多次反复进行模型调试，甚至需要对原始数据进行二次清洗和重新分箱。

（1）VIF

```
vif(fit.reduced)
```

方差膨胀因子（Variance Inflation Factor，VIF）：容忍度的倒数，VIF 越大，显示共线性越严重。经验判断方法表明：当 0<VIF<10，不存在多重共线性；当 10≤VIF<100，存在较强的多重共线性；当 VIF≥100，存在严重多重共线性。

（2）预测

```
    Pdata=predict(fit.reduced,newdata=testdata)#预测在该模
型下测试数据的表现
```

（3）ROC 曲线

ROC 曲线是用来确定判定为 0/1 的分割点位置：

```
    TPR=rep(0,1000)
    FPR=rep(0,1000)
```

```
for (i in 1:1000){
+     p0=i/1000
+     ST=d
+     SP=1*(p>p0)
+    TPR[i]=sum(ST*SP)/sum(ST)
+    FPR[i]=sum((1-ST)*SP)/sum(1-ST)
+ }
plot(FPR,TPR,type='l')
```

效果如图 7-42 所示。

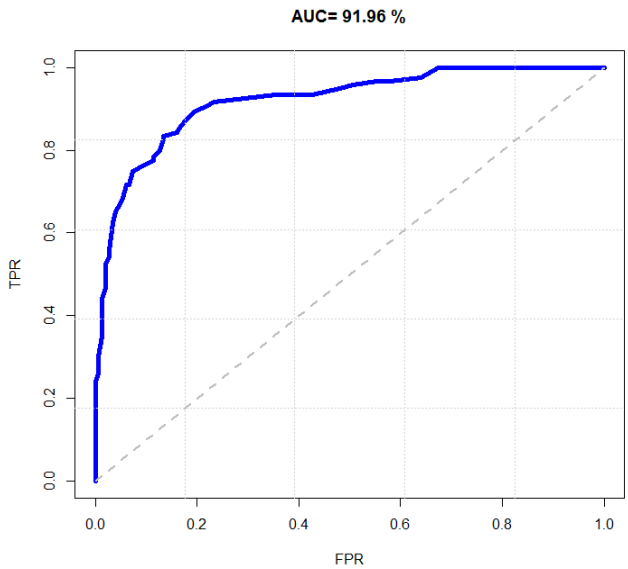



图 7-42

（4）K-S 值计算

K-S 值是用来衡量模型区分度的重要标准，logistic 回归模型一般要求 K-S 值在 30%以上才算做有效。

```
AC=testdata$due
KSDATA=data.frame(p,AC)
ORDERDATA=KSDATA[order(KSDATA$p),]
AC0=sum(ORDERDATA$AC==0)
AC1=sum(ORDERDATA$AC==1)
```

```
        for(i in 1:nrow(ORDERDATA)){
        ORDERDATA$PDSUM0[i] = sum(ORDERDATA$AC[1:i]==0)
        ORDERDATA$PDRATE0[i] = ORDERDATA$PDSUM0[i]/AC0
        ORDERDATA$PDSUM1[i] = sum(ORDERDATA$AC[1:i]==1)
        ORDERDATA$PDRATE1[i] = ORDERDATA$PDSUM1[i]/AC1
        ORDERDATA$KS[i]=abs(ORDERDATA$PDRATE0[i]-
ORDERDATA$PDRATE1[i])
      }
    K-S=round(max(ORDERDATA$KS),4)
```

**小提示：提供一种增大样本量的方法**

从样本中多次随机选择记录组成一个新的数据组。

```
rr=NULL
for (i in 1:100){
+    ri=alldata[round(runif(10,1,311)),]
+    rr=rbind(rr,ri)
+ }
```

（5）模型使用

最终我们得到公式：

$$Y=\ln(p/(1-p))=b_0+b_1\times x_1+b_2\times x_2+...+b_n\times x_n$$

通过这样的公式产生的 $Y$ 值为概率值，也即预测结果为 1 的概率，在传统评分卡制作中，我们把用 $1000-Y\times1000$ 作为用户的信用分数。

至此，逻辑回归案例就结束了，大家不妨仿照这个案例自己手动尝试做一个模型看看实际效果如何，加深对建模的印象。

# 第8章 初识R语言

我们不需要 Print Hello World；

我们不需要了解向量、矩阵、数组、数据框的区别；

我们不需要了解 For、While、Switch；

我们直接开始工作。

## 8.1 安装与编辑器

R 语言由于其开源免费的属性，安装起来极其简单，不需要购买、付费、破解等繁琐的流程，同时 R 语言大小不到 100MB，在所有数据分析类软件中算是极小的了，使其下载安装起来极其方便，大家可以直接搜索 R 就可以找到开源社区 https:// www.r-project.org。

在网站中找到合适自己系统的版本，通过“download R”选择自己所在的国家与大学提供的下载渠道，下载后直接安装。

R 语言安装完成之后打开的页面如图 8-1 所示。

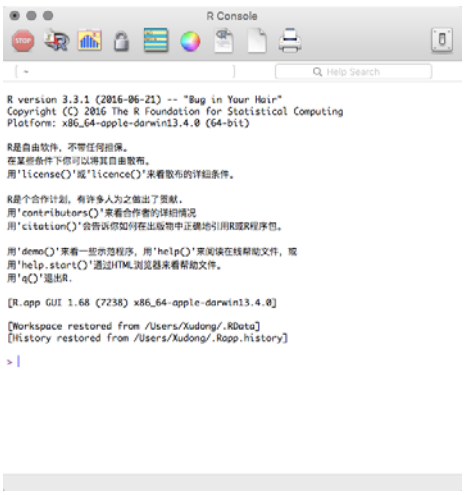



图 8-1

由于其按［Enter］键立即执行的功能并不适合直接在这个窗口中编程，就像许多编程语言一样，我们需要安装一个编辑器，这里推荐 RStudio，同样是完全免费而且体量很小的编辑器，函数提示和括号自动补充等功能都十分便利，如图 8-2 所示。

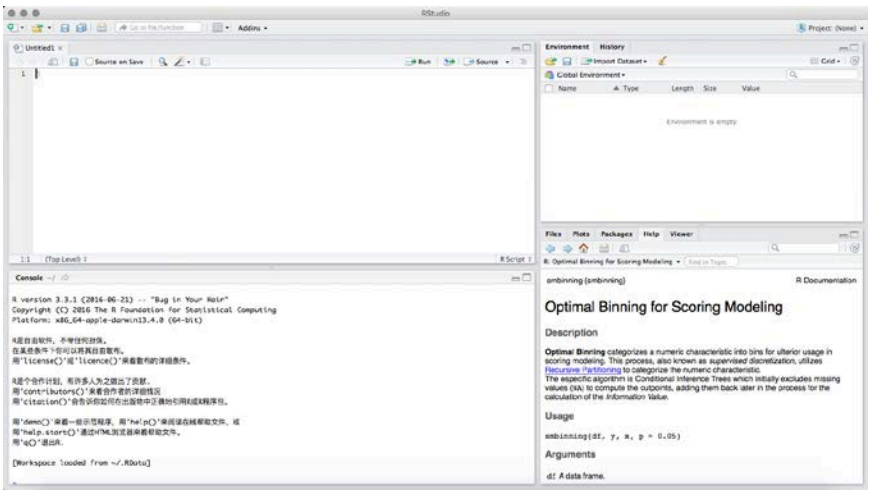



图 8-2

大家可以通过新建一个 Rscript 直接开始编程。

安装完 R 和 RStudio 之后，我们还会常常遇到的问题就是 R 语言的包，R 语言有许多包能够辅助进行数据处理，许多复杂的编程算法在 R 语言里有些包可以通过一个函数就能完成，极大地简化了工作量。R 语言包的安装是直接通过指令实现的，不需要单独去网站寻找，虽然是一个简单的问题，但许多时候还是会对新手造成一些困扰。

安装 R 语言包使用如下语句就可以实现：

```
>install.packages(“package_name”)
```

双引号中放入你想要安装的包的名称，例如：

```
>install.packages(“vcd”)
```

上述语句我们就安装了一个叫作 vcd 的用于可视化类别数据的包。

安装完成之后，如果下次再使用这个包需要使用语言：

```
>library(vcd)
```

因为随着时间的积累，你需要下载很多 R 语言的包，如果 R 语言每次启动都把这些包全部加载势必占用许多内存，因此，每次使用包的时候需要手动调取这个包。

大家如果想要详细了解每个包的内容或者是 R 语言使用中遇到任何问题都何以使用：

```
>help(vcd)
```

也可以使用下属语句来查询帮助文档：

```
>??vcd
```

R 语言的帮助文档都是英文的，好的英语水平在软件学习中会很有帮助。

## 8.2 数据读取

在安装 R 语言之后其实不需要先了解常用编程命令和 a=1 等基础功能，许多东西在实践中自然而然就学会了，为此单独耗费时间着实没有必要，让我们直接开始干活吧。

数据分析的对象要么是你手头已经有的数据，要么是数据库里还未导出的数据。由于 Excel 格式是微软各类算法集成的庞大表格，建议大家都把数据集保存成 csv 的格式，方便 R 语言的读取和数据处理。

假设我们有一个 data.csv 的数据集，把它放在桌面上，使用 R 语言的读取方式如下：

```
>datasample=read.csv("/Users/User/Desktop/data.csv",
header=T)
```

这样，我们就把数据读取进入 data 数据集中，在 R 语言中 data 叫作数据框，就像之前所讲，我们不用区分向量、矩阵、数组、数据框，我们在使用 R 语言进行数据分析时大多面对的是这种数据框，我们先熟练运用数据框的处理方法。

想要知道这个数据框的内容直接输入：

```
>datasample
```

你就可以看到这个数据集的全貌了。然而许多时候由于数据量极大，如果想大致了解下这个 data 数据框的数据结构，则可以使用：

```
>summary(datasample)
```

可以看到数据框中每个变量的平均值、最大值、最小值等。如果你想看这个数据框的某一个元素，则可以使用：

```
>datasample[2,3]
```

意在查看数据框的第 2 行第 3 列的元素；如果想看一个范围内的元素，则可以使用：

```
>datasample[1:10,1:10]
```

可以看到第 1 到 10 行的第 1 到 10 列的数据；如果想看 1 到 10 行的所有数据，则可以使用：

```
> datasample[1:10,]
```

同理，想要看 1 到 10 列的所有数据，可以使用：

```
> datasample[,1:10]
```

如果想看数据框的所有变量名称，则可以使用

```
>names(datasample)
```

知道数据框的变量名称之后，如果想看某一个变量的值，则可以使用：

```
>datasample$var
```

上面的 $ 表示引用这个数据框的 var 变量，我们还可以通过：

```
>table(datasample$var)
```

查看这个变量的频度分布。

截止这里我们基本上了解了 R 语言是怎么读取数据的以及如何查看所读取的数据，读取 csv 格式的数据基本上是数据分析前期需要用到的唯一方式，而数据框的各种查看方法是帮助我们在进行数据处理时能够在每一步校验输出结果，根据不同的需求进行数据观测。

## 8.3 数据处理

如果说数据分析是做菜，了解了数据读取之后就好像找到了要做的菜，接下来就是如何处理它们的事情了。

数据处理简单来说无外乎：加、减、乘、除；增、删、改、查。

在 R 语言中，它们的算法与 Excel 中几乎没有差别，唯一需要注意的就是变量的数据类型。

假设 datasample 数据框的内容如表 8-1 所示。

表 8-1

| date | var1 | var2 | var3 | var4 |
|---|---|---|---|---|
| 1 月 1 日 | 1 | 56 | 34 | K2 |
| 1 月 2 日 | 2 | 73 | 77 | K5 |
| 1 月 3 日 | 1 | 35 | 93 | K7 |

如果我们输入：

```
>datasample$var1 + datasample$var4
```

系统将会报错，因为 var4 属于字符型，而 var1 属于整数型，我

们想把 var2 与 var3 加起来做一个 var5，语句如下：

```
> datasample$var5 = datasample$var2 + datasample$var3
```

语句中在“ = ”前面的 datasample$var5，表示在 datasample 中新建一个变量 var5 将 datasample$var2 + datasample$var3 的值赋给它。

在对变量进行处理的过程中，我们要关注日期变量，R 语言常常会以字符型的格式读取日期变量，只需要用一个函数修改下就好了：

```
> datasample$date=as.Date(datasample$date)
```

如此，我们通过 as.Date 函数把字符型转换成了日期型。

在加减乘除的过程中，除法常常会伴随小数点，设置小数点位数的方法有两种，一种是在开头声明小数点的位数：

```
>options(digits=2)
```

另一种是单独对变量进行处理：

```
>round(var,2)
```

两种方式都可以用来设置小数点的个数。

我们再来看看增删改查中的：删。最简单的方式如下：

```
>datasample1=datasample[,-2]
```

这样我们就把 datasample 的第二列删掉了，如果我们想删除多列可以执行如下指令：

```
>datasample1=datasample[,c(-2,-3)]
```

这样我们就把第二列和第三列都删除了。

再说说改，我们可以直接把数据框调出来手动修改：

```
>fix(datasample)
```

也可以通过 reshape 包：

```
>library(reshape)
>newdata=rename(datasample,c(var1="newvar"))
```

这样就完成对变量名称的修改。

至此，我们对数据进行处理时遇到的加、减、乘、除和增、删、改、查都基本了解了，在实际工作中还会遇到许许多多的数据处理的问题，我的建议是：

遇到问题，解决问题!

只有结合实践的学习才是深刻的学习。

## 8.4 经典算法

在数据挖掘行业有十大经典算法，许多项目都可以进行拆分到这些算法上进行解决，这些算法在 R 语言中都有实现的案例，十大算法如下：

（1）Apriori：是一种最有影响的挖掘布尔关联规则频繁项集的算法。

（2）C4.5：是机器学习算法中的一种分类决策树算法，其核心算法是 ID3 算法。

（3）Naive Bayes：在众多分类方法中，应用最广泛的有决策树模型和朴素贝叶斯（Naive Bayes）。

（4）K-means 算法：是一种聚类算法。

（5）SVM：一种监督式学习方法，广泛应用于统计分类以及回归分析中。

（6）CART：分类与回归树，下面有两个关键的思想，第一个是关于递归地划分自变量空间的想法，第二个是用验证数据进行减枝。

（7）KNN：是一个理论上比较成熟的方法，也是最简单的机器

学习方法之一。

（8）Pagerank：是 google 算法的重要内容。

（9）adaboost：是一种迭代算法，其核心思想是针对同一个训练集训练不同的分类器然后把弱分类器集合起来，构成一个更强的最终分类器。

（10）EM：最大期望值法。

这十大算法，R 语言都有对应的逻辑包进行统计和处理，下面就以 Apriori 算法为例，我们使用 R 语言中的 Matrix 包进行数据挖掘算法。Apriori 算法的最早提出是为了寻找关联规则，因为其有很清晰和简单的算法逻辑结构，所以逐步成为一种搜索算法思想，有点类似于动态规划，贪心算法的概念。

```
>install.packages("Matrix")#安装 Matirix 程序包
>library(arules) #加载 arules 程序包
>shop=read.csv("shopping.csv",header=T)
>data(list=shop) #调用数据文件
>mode(shop)#确定数据集是可被用于进行关联规则的类型
>shopnew=shop[8:18]#选取进行关联规则的变量
>rules=apriori(shopnew,parameter=list(support=0.1,co
nfidence=0.8)) #设定最小支持度为 0.1,最小规则置信度为 0.8,求此条
件下的关联规则
>frequentsets=eclat(shopnew,parameter=list(support=0
.1,maxlen=800)) #求频繁项集
>inspect(frequentsets[1:10]) #观察求得的频繁项集
>inspect(sort(frequentsets)[1:10])#根据支持度对求得的频
繁项集排序并察看 #
```

最后我们得到如下结果：

```
items      support
#1  {蔬菜制品}   0.303
#2  {冷冻食品}   0.302
#3  {果蔬}       0.299
#4  {啤酒}       0.293
#5  {鱼类}       0.292
#6  {红酒}       0.287
#7  {糖果}       0.276
```

```
#8  {肉制品}     0.204
#9  {软饮料}     0.184
 #10 {鲜肉}       0.183
```

我们的结论是蔬菜制品和冷冻食品（支持度为 0.173），冷冻食品和啤酒（支持度为 0.17）以及蔬菜制品与啤酒（支持度为 0.167）是最可能连带销售的食品。

有兴趣的读者可以研究研究其他的算法在 R 语言中的实现，它们大多是以 R 语言包的形式实现。

# 第9章 行业的未来

## 9.1 市场需求

通过之前的章节我们了解了大数据时代的现状以及数据分析师需要做的主要工作，接下来我们简单地了解下数据分析师的职业发展。

其实了解一个职业的发展情况可以从各个公司的 HR 招人时给的奖励来看，一般给的奖励越多这个岗位也就愈加稀缺和重要。身边有许多认识的 HR 让我帮忙推荐数据分析师。成功推荐一个参与面试的数据分析人员，且不论这个人最终能否被录取就能得到几百元乃至上千元的报酬，这样高的福利标准充分地反映了市场上对数据分析师的需求多么急迫和不遗余力。

随着互联网行业的发展，现在几乎所有的公司在开展业务的过程中都会或多或少地借助互联网的力量，而借助互联网开展工作就一定会有底层数据库和数据结构，这些通过数据库沉淀的数据仿佛是企业的一座座未知的金矿，亟待专业的人才帮忙挖掘其中的价值。如果说互联网行业近十年来在中国呈现爆炸式的发展，那么相应的互联网人才难免会跟不上步伐。计算机相关专业因为是互联网发展的刚性需求发展得较为迅猛，起步也相对较早。而数据分析行业由于其在行业初期的重要性不是很突出一直被大家忽视，随着互联网行业竞争进入红海，越来越多的公司开始关注数据分析来挖掘前期爆发式增长所带来的剩余价值，这也是互联网行业这两年才出现的爆发式增长的需求。与此同时，数据分析行业在各类大学中并没有相关专业，数学和统计学作为相对冷门的学科在之前并没有得到大家的重视，也就造成了现在人才缺失的局面。一方面是旺盛的市场需求，另一方面是从业人员相对较少，造成了现在的供需极度不平衡。

市场上对数据分析师的一些要求大体上有如下几点，如图 9-1 所示。

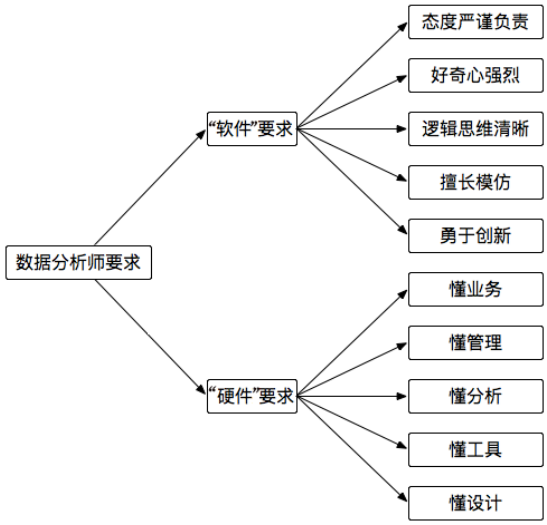



图 9-1

（1）“软件”要求

态度严谨负责即要求一名合格的数据分析师应具有严谨、负责的态度，保持中立的立场，客观评价企业发展过程中存在的问题，为决策层提供有效的参考依据；

好奇心强烈指的是作为数据分析师，要积极主动地发现和挖掘隐藏在数据内部的真相；

逻辑思维清晰要求数据分析师具备缜密的思维和清晰的逻辑推理能力；

擅长模仿是能够领会他人方法的精髓，理解其分析原理，透过表面直达本质；

勇于创新是一个优秀分析师应该具备的精神，使自己站在更高的角度来分析问题，为整个研究领域乃至社会带来更多的价值。

（2）“硬件”要求

懂业务是要求数据分析师能够熟悉行业、公司业务及流程，能

够有自己独到的见解；

懂管理一方面是为了确定分析思路、搭建数据分析框架，另一方面的作用是针对数据分析的结论剔除没有指导意义的分析决策；

懂分析是指掌握数据分析的基本原理与一些有效的数据分析方法，并能够灵活运用到实际工作中，以便有效地开展数据分析工作；

懂工具是指掌握数据分析相关的常用工具，数据分析的方法是理论，而数据分析工具就是实现数据分析方法理论的工具；

懂设计是能够运用图表有效地表达数据分析师的分析观点，使得分析结果能够一目了然。

由于市场上数据分析师的普遍欠缺，现在有许多专门做数据分析的公司为没有数据分析师的企业提供数据服务。数据分析师为本公司提供服务和为其他企业提供数据分析服务的职能有所不同，各有特点，同时近些年也在不断发展变化。为其他企业提供数据分析服务主要是以一个一个项目的形式，提供数据分析报告，建立数据模型，推荐策略方案。这决定了在为其他企业提供数据分析服务的分析师可以快速接触不同企业的项目，学习相对完整的分析流程，也可以快速结识行业内部不同的人，看看各个企业都有哪些需求，开阔眼界。因为服务的企业分布不同，可能需要经常出差，这对于年轻人来讲不是问题，对于有家庭的人可能是个需要考虑的问题。

为本公司提供数据分析服务的数据分析师最早的时候主要负责项目的企划和实施的管理，不自己做具体的数据分析，但后来很多企业开始建立自己的分析团队，算是数据分析师本地化、专业化的一个过程。Capital One 在美国零售金融业中，数据分析应用是数一数二的，但它当年最初也由 Experian 等专业公司提供数据建模服务，后来慢慢成立自己的数据团队，和业务整合越来越深。为什么公司要自己建立分析团队，主要是为了将分析和业务更好地整合。数据分析工作的价值最终要靠改善业务决策体现出来，要做得好，就要更了解业务。相对于在为其他公司提供数据分析服务的数据分析师

转移到只针对一家公司提供服务后，数据分析师可以更好地了解自己企业的业务，比如在做数据分析之前，先轮岗到业务部门锻炼一段时间，了解实际流程。更重要的是，在做了分析之后，可以持续和业务部门合作，将分析结果应用到业务中去，了解实际结果是好是坏，反馈改进。这一点，做外包项目的数据分析师可能比较难做到，因为模型建完，报告递交后，项目就结束了，你不知道别人到底如何用的模型，结果到底怎么样。

这对于企业自身的数据分析师是动力也是压力，最后结果应用效果不好，你可是跑不了的，不是交了报告就行。所以最后倒逼分析师要不断参与整个流程工作，从数据的生成、搜集，到最后的实施过程都要参与负责，从以前只是提交报告“顺利地做完”，转换到“曲折地做成”，中间会有不小的转换。

从数据分析服务公司来看，近几年数据分析服务有了新的变化，国内更多数据分析方面的创业公司采用 SaaS 化的方式，商业模式也从基于项目式转换到基于产品，甚至是基于交易的形式，并且提供从数据采集、数据分析和决策跟踪整个流程的服务。在这些新型公司中，数据分析师也要分析自己公司的客户行为，作为 SaaS 型的分析平台，客户量往往比过去项目型时多得多，每个客户可以使用各种功能，同时也可以随时退订相关的功能，取消付费，这就存在拉新、激活、推荐、挽留等各种客户运营问题，存在各种客户分析的应用场景，分析师要帮助产品经理、运营人员不断优化产品流程，这本身就是一个公司自身分析师的职责。另外由于是全流程的平台，所以分析、决策的结果都能留存在这个平台上，作为为客户服务的乙方分析师，也可以不断实践和改进自己的工作，同时也可以接触不同的用户，了解不同用户的需求。

从构建自身数据分析体系的公司来看，在很多企业，比如比较大的互联网公司，数据工作逐渐平台化，在整个企业内成为最基础的服务平台，通过建立各种数据产品，来服务各个业务部门，这个过程就需要向提供数据分析服务的公司一样，不断在内部做营销，

推广数据理念，让业务部门把数据用起来，这样数据部门才能争取更多的资源。这时就要求数据分析师把业务人员当客户，既要有服务的心态，又要发挥沟通、影响的作用。

随着时代的发展，一个人一生的职业生涯中可能会不断转换角色，可以是为自己公司搭建数据分析体系，也会为其他公司提供数据分析服务，唯一不变的应该是创造你自己独特的价值，建立人脉网络，树立个人品牌。无论以怎样的身份进行数据分析服务，作为一个数据分析师，这三件事都是可以做的。

（1）干好手头的工作：建立个人品牌口碑最重要，在各个岗位都能及时高质量地完成工作，未来就有更多机会，职业圈子不大，当有人想找人的时候，总是先想到最靠谱的人。

（2）建立作品库：可以写文章，也可以在 gitpub 或者 rpub 上建立自己的库，你的作品是你最好的简历，这个过程不但能让自己及时总结，不断提高，也可以帮助别人，提供价值，在相互帮助中建立人脉网络。

（3）沟通交流：互联网时代，交流的方式越来越多，微信群或者线下沙龙都有很多，作为听众可以多吸收其他同行同业的经验，作为分享者也可以梳理自己的思路，提高自己的能力。

作为一名数据分析师，你随时都是在为人服务，为别人提供价值，同时也从别人那里得到帮助，未来的企业会越来越成为一个平台，是自由人的自由联合，在其中经营好自己，同时也给他人带来价值。

## 9.2 重要性、必要性

市场需求来源于各个企业对这个岗位的重视和人才的缺失，另一方面也反映了这个岗位对公司的重要性和必要性。我们一直在说数据驱动，数据分析可以说是企业的大脑，在数据分析的过程中发现公司的问题，给出解决方案和优化建议，监控日常业务的开展和寻找业务中的亮点和闪光点加以发扬光大。

在公司运作的过程中，不仅是数据分析人员需要专业的数据分析能力，各团队负责人、一线工作人员、开发工程师都需要对数据分析工作有一个清晰的认知和了解（图 9-2）。

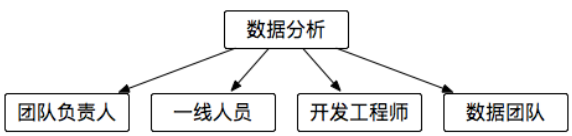



图 9-2

各团队负责人很多时候是数据的消费者，首先是要学会看报表。比如 Vintage 图、留存率图这些报表，一开始不是特别容易理解，要细致地一点一点讲明白，还要看出意义。看报表的核心是做细分，然后做对比。对比不同行业、不同地区、不同时间、企业内外，找出差距、定位问题，当数据的消费者有更清晰的思路时，就可以提出更加合理的需求，即节省数据团队的时间，也提升自己团队的效率。各个团队负责人最重要的工作就是做各种决策，包括方向性的决策和操作性的决策。相比传统成熟行业，创业公司的决策场景，会有更多的不确定性，创业就是在试错，这就需要了解一些风险决策的思想和方法。每个决策都可能有不同的结果，判断一个决策好不好，不是只看最后结果，就像买了保险，最后没有出险赔偿，不能说买保险这个决策是错误的，很多时候未来是不能预测，只要在平均期望上达到最佳就可以，不能以成败论英雄，这个主题可以另

开一篇，更加深入地探讨一下。给业务人员讲概率、风险、决策的内容，不能讲得太理论，就像加州大学戴维斯分校的蔡知令老师所说：统计知识要讲得让祖母也听得懂，才能影响更多的人，才真的有用。他在戴维斯商学院给 MBA 学生讲授统计课，被 MBA 学生 14 次评选为年度教师，除了学术严谨，更主要的原因是学生能够听懂，并且用在自己的工作中。这方面还需要不断地探索和总结，一个例子对于知识点需要做适当的简化，但又能体现出核心的思想，这是不容易的，前辈看似信手拈来的例子，其实背后也是几经挑选和打磨，才能拿出来分享的。

一线人员包括财务、客服、运营等各个部门都需要对数据分析有一定的认知。对于创业公司来说，可能还没有非常完善的后台系统，很多时候需要用 Excel 完成很多工作，这时候学习一些基本的 Excel 技巧，就能大大提高工作效率，同时也减少人员的流失。这个内容里面包括一些主要的函数，比如 vlookup、match 等，还有相对引用的公式，透视图，透视表等，网上有很多 Excel 课程，但是人们往往没有毅力学下去，或者看了之后在工作中用不起来。所以这种培训不能只是讲 Excel 功能，而是要对照实际工作流程，找到典型的重复工作场景，实际案例，再结合 Excel 功能来讲。只有在每日面对的繁重工作瞬间完成的那一刹那，人们才能真正体会这些课程的作用。

公司开发工程师是另外一个需要对数据分析有认知和了解的群体。开发工程师使用 Python 做开发，数据团队的分析师也是用 Python 做数据处理和建模，在工具上没有障碍，数据团队相互学习分享的时候，也会吸引开发的工程师一起交流，一方面开发工程师了解基本数据概念和方法后，在某些后台功能的开发时，可以和数据团队更好地衔接，对于非常有兴趣，深入钻研的人员，也会有机会转到数据团队来工作。这个方面的内容包括基本的统计概率知识，比如不同的分布、均值、方差、估计等，这里推荐一本参考书，*Think Stats : Probability and Statistics for Programmers*，以 Python 为工具来讲解统计学的基础知识，作者还有一系列的相关书籍，都是以 Python 为

工具，比较推荐。另外也会有机器学习相关的内容，包括 Python 的 scikit-learn 库及其相关概念的介绍，scikit-learn 库的帮助文档非常好，不仅有库函数的介绍，还有机器学习相关算法的介绍，是个很好的入门教材。

作为核心和关键的数据分析团队的人员对数据分析必须要有强烈的把控力和专业度。数据团队内部的培训更多是教学相长的方式，每个人都要自学，自己尝试实践，然后准备自己的主题，把学习的结果和经验贡献给其他人。这种方式不仅提高了团队整体的学习效率，也能改善主讲人个人的学习效果，从学习金字塔可以看出，学习内容留存率最高的就是教授给他人，这也是教学相长的一个体现。

数据相关的课程网上有很多，但其中最难讲，也比较少讲的就是数据诊断清理。斯坦福统计教授 David Donoho 于 2015 年在他的文章《数据科学 50 年》中，也提到了这个问题。诊断、清理、整合数据在数据分析工作中占到 70%以上时间，对于结果的影响很多时候也超过模型的选择，但是在实际数据课程中却比较少提及。这其中原因包括“教”和“学”两个方面，一方面是这个工作有更多经验性的内容，不像讲模型算法那么清晰明了，不好讲；另一方面这些工作都是平时所说的脏活累活，不像建模算法那么高大上，刚入门的同学反而不愿意听这部分内容。可能在工作中会对于不同数据工具这部分有针对性内的容进行练习，SAS 中就是 data 步的工作，在 R 里有 reshape、dplyr 包，在 Python 里有 pandas 包，在 Spark 里也有 spark sql 模块，熟练运用这些工具，把数据像削瓜切菜一样，整成不同的丁丁块块，才能准备好进入下一个步骤分析建模。

这部分如果是做成课程，那最好不要每个命令用一个单独的数据集，而是使用一个完整的数据集，针对一个建模目标、模拟实际情况，覆盖主要的数据处理命令和函数，这样练习的人会更有实战的体会，更接近实际的需要。创业企业变化快，情况各异，大家都可以尝试和寻找适合自己企业的数据课程的内容和方法，但最终目的还是让不同人都能体会数据的好处，不求高深，只求对工作有用，

让“数据”这个词更加深入人心。

企业在各个方面都需要数据分析来辅助进行决策（图 9-3）。

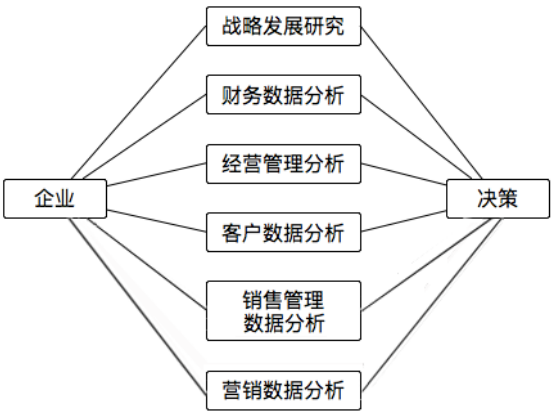



图 9-3

战略发展研究对于企业来说，在竞争激烈的市场环境下，建立科学合理的企业战略是确保企业良性循环发展的重要保证。企业战略发展研究，为企业战略的制定与实施提供了重要的参考价值，企业通过定期的企业战略发展研究，可以在企业战略管理根据环境适时加以调整的过程中做出重要贡献。

财务数据分析是对财务数据进行分析，可以帮助企业中财务报表的使用者或管理层能更好地了解、掌握一个企业的生产经营情况。及时准确地对企业财务数据进行分析，具有重大意义，其表现如下：

第一，通过分析资产负债表，可以了解公司的财务状况，对公司的偿债能力、资本结构是否合理、流动资金充足性等做出判断。

第二，通过分析损益表，可以了解分析公司的盈利能力、盈利状况、经营效率，对公司在行业中的竞争地位、持续发展能力做出判断。

第三，通过分析现金流量表，可以了解和评价公司获取现金和

现金等价物的能力，并据以预测公司未来现金流量。

经营管理数据分析指的是伴随着市场竞争的日趋白热化，对于经营管理数据分析作为企业提升自身经营管理水平的重要内容被愈来愈热点关注。在信息技术急速进步的背景下，对经营管理数据的分析已然成为评价一个企业经济管理水平的一项重要参考依据。企业经营管理数据分析是调查研究不可缺少的重要环节，是充分发挥企业调查研究在经营管理中的重要保证。

市场数据分析不仅可以研究产品及服务购买者及消费用户的心理和行为，而且可以对市场营销活动的所有阶段加以研究，通过对从生产者到消费者这一过程中的全部商业活动的资料和数据做系统的收集、记录、整理和分析，对于企业了解和发现商品的现实市场和潜在市场具有重要意义。

对客户数据分析是有关客户评价数据的研究，其内容非常丰富，以客户满意度分析为主要标志。客户满意度研究近年来在国内外得到了普遍重视，特别是服务性行业的客户满意度调查已经成为企业发现问题、改进服务的重要手段之一。所以通过客户数据分析了解客户的需求、企业存在的问题以及与竞争对手之间的差异，从而有针对性地改进服务工作，显得尤为重要。

销售管理数据分析是对于企业在经营中产生大量的与销售有关的数据进行收集、整理、分析、存档。对销售管理数据进行分析可以很好地辅助企业制订销售计划与决策，帮助企业快速、准确地了解执行结果，提高营销系统运行效率。

加强营销数据的采集与管理并进行合理、准确、有效的实时性分析，有助于企业逐渐克服经验营销导致的局限性或对经验营销者的过度依赖性，形成科学的营销新理念，提高企业的市场认识能力、市场管理能力和市场适应能力。

在大型的传统公司，数据分析又有不同的模块和应用场景（图 9-4）。

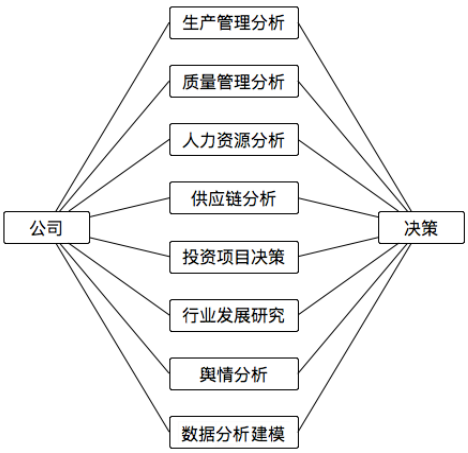



图 9-4

生产管理是企业竞争的关键因素，提高设备、提高人员工作效率、提高质量管理效率，节约资金、节约劳动力、降低人为因素，是企业生产管理自我改善的根本目标，生产管理数据分析，对于企业高效、低耗、灵活、准时地生产合格产品，以及保证实现企业的经营目标、有效利用企业的制造资源，对适应市场、环境的迅速变化等具有重大价值。

企业质量管理数据分析是指组织应确定、收集和分析适当的数据，以证实质量管理体系的适宜性和有效性，并评价在何处可以持续改进质量管理体系的有效性。企业质量管理数据分析是精确分析质量管理现状的科学方法，它是质量管理实现持续改进的有效途径和重要手段。

随着人力资源管理理论和管理实践的迅速发展，人力资源管理的各大模块的职能已趋完善，如何提升人力资源管理的价值，是传统的人力资源转型的核心，在这当中，人力资源管理的专业化水平的提升是人力资源管理职能扩大和深化的关键。而在人力资源专业

化的提升过程当中，人力资源管理数据分析扮演了至关重要的角色，它使得人力资源管理的理念、技术及技巧在定量方法的分析下，可以趋于合理化，更加的科学化。

目前在采购和供应链上应用大数据的重心更多是靠近市场的需求端和营销领域，相对于采购与供应领域，市场需求领域更多地首先开展了大数据的应用，许多企业也已经收获颇丰。因此，在采购与供应领域应该努力迎头赶上时代的步伐，利用大数据为企业和供应链的供应做出更大的贡献。

投资项目决策是企业管理中最重要的决策，对企业的获利能力、资金结构、偿债能力以及长远发展都有着直接影响。随着我国市场经济的发展，市场竞争日益激烈，投资主体和投资渠道趋于多元化发展态势，如何优化配置资源，有效的利用资源，提高企业投资决策水平和效益，是当前企业经营发展中的突出问题。因此，能否做出正确的资本决策是工程项目经营发展的关键。

行业是由许多同类企业构成的群体。如果我们只进行企业分析，虽然我们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，则无法通过比较知道目前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。

只有进行行业分析，我们才能更加明确地知道某个行业的发展状况，以及它所处的行业生命周期的位置，并据此做出正确的投资决策。

随着网络的发展，舆情对企业的影响越来越大，建立舆情监测系统对企业发展越来越重要，企业需要加强舆情监测工作。目前网络舆情主要来自于各类新闻、微博、微信、博客、论坛贴吧、视频网站等网络平台，对于企业来说，要根据地理位置、行业、影响力范围等企业自身的特点对舆情监测源进行综合评估，以确定市场最新动向、消费趋势，以及市场需求动向。因此，舆情分析对企业的

成长和发展具有积极意义。

数据分析建模是一种量化的思考方法，是运用数学和计算机的语言和方法，通过抽象，简化建立能近似刻画并“解决”实际问题的一种强有力的数学手段。

随着互联网的发展，数据急剧累积，纷繁复杂的社会经济活动表象越来越难以让人看清背后的实像，而数据模型可以从市场行为产生的繁多现象背后抽象出数量关系，最大可能消除项目中人为的主观臆测和判断。随着大数据时代的到来，数据模型分析显得越来越重要。

## 9.3 大数据，下一个风口

大数据交易市场，到底在交易什么？

**一切为了用户所思即所得！**

**一切为了市场所给即所要！**

用户想要什么我们就给他什么，我们给用户什么用户就想要什么！一句话：我们要知道用户的需求在哪里。怎么知道？通过市场调研还是数据分析？市场调研的手段有很多，一种常规的手段就是问卷调查，而大数据可以说是一种可以替代问卷调查并极大提高调查效果的一种手段。问卷调查的核心思想是抽样调查，而抽样调查的工作方式是通过少量随机样本反映整体数据表现的一种方法。其特点是：选取少量样本，调查目标关键特征。大数据做的第一步是把少量样本变成所有样本，这样就完全排除抽样调查中样本不够随机、不够均匀的问题。大数据做的第二步是把调查的关键特征扩大到所有可能的特征，解决了问卷调查中问题设计等信息获取不全面的问题。从这个角度来说大数据完全可以在市场调研中发挥极大的作用。

哪来的数据？数据交易。长久以来数据分析一直是对已有数据进行分析，挖掘用户的喜好、分析用户的行为。然而第一批用户总是最难获得的，同时第一批用户的质量和特征极有可能是由我们的推广方法决定的。一开始采用某一方法进行推广，带来了对应的用户，我们对这一些人进行数据分析，挖掘他们的行为偏好极有可能是不客观的，得到的结论也会有所偏差。这个时候我们就想：能不能在数据源上打开窗口？

数据源交易的是什么？

**市场行情、用户习惯、用户信息**

上策交易市场行情，中策交易用户行为数据，下策交易用户的具体信息。

在互联网行业疯狂滋长的年代很容易让人想到中国民营企业急速发展的那一段岁月，在野蛮生长的那一段时间，很多民营企业的原罪并没有得到及时有效的遏制，人们放纵欲望的滋长，通过各种渠道谋取利益，产生了很多急速扩张的暴发户，但是当泡沫破灭之后很多人因此破产甚至进入监狱。这些很容易让人联想到在互联网急速扩张的这几年，互联网行业有哪些原罪？

当下正逢互联网急速发展的时代，身处浪潮之巅的人们很难静下心来去思考我们的时代到底有哪些原罪，即使思考也难免片面和偏激，“只缘身在此山中”的状态让一切都很难讲清楚，只是希望聊胜于无，在不远的将来，当互联网泡沫破灭的那一天，好有个心理准备，不要等到泡沫破灭才相信有泡沫。当然这句话是有争议的，如果你反驳我。说现在互联网没有泡沫，或者现在的泡沫是“好的泡沫”，起到了加速和催化作用，一切正在朝着好的方向发展。我甚至可以赞同你，但不妨未雨绸缪，思考一下泡沫这件事，那我们来谈谈互联网企业的原罪。

作为一个互联网数据行业的从业人员，我深知数据对一个公司的重要性，对内这是衡量一个企业发展的绝对标准，你不能说员工

都很有干劲就说明正在做的事情很靠谱，你需要知道业务的数据表现、发展趋势、营收情况等。对外数据是别人了解一家公司的首要途径，你有多少用户，你有多少市场……这些都是衡量一个企业未来的重要参数，但是现在有多少互联网企业在做虚假的数据？对外欺骗来获得投资和市场，对内欺骗来获得效率和忠诚。由于网络传播的及时高效，让人们吹的“牛”一夜之间就能火遍全中国，这里不好点名，但是作为一名互联网从业人员，看看四周，你很容易就能看出有几个吹牛吹得收不回来的，又有多少是看不出来的呢？不得不说当下是一个靠着吹牛就能挣钱的时代，你说那么多人靠着吹牛拿了那么多钱，我为什么要诚恳？是啊，所以说这是互联网企业的原罪。

互联网颠覆一切的冲动和盲目，使人们想要颠覆任何行业，让一切联网！餐饮娱乐、衣食住行，哪一项没有奔跑着拥抱互联网，连出版社都要被颠覆，任何人都可以出书，做自出版，自己做媒体公关，自有一大帮人过来买单，不远的将来，选择一本好书对个人的要求真的太高了！网上有多少假货？你网上买了多少“一次性用品”？实体衣服店真的被取代了吗？人们真的像想象中的“在实体店试衣服，到网上买”？人们用互联网这一根棍子捅向一切可以接触的行业，让它与时代接轨，跟得上移动互联，谁知道这是不是破坏与捣乱呢？冲动、粗鲁、野蛮的想要颠覆一切，人人都觉得颠覆一个行业能挣到很多钱，我要做第一个。

你有梦想吗？你想发财吗？你想大干一番吗？跟我干，有前途！虽然现在艰苦了点，但是我们有未来。你看行业现状、身边案例，你看我们的数据……很多时候我们都没有能让我们双脚坚实地站在大地上的东西。什么是坚实的大地？我们的公司能挣钱。而我们现在的状态是：有了用户，盈利模式自然就来了。我不对这种模式做好或是坏的评价，我只是客观地表述这种企业模式毕竟有些虚浮，未来的不确定因素太多，谁又能料到未来呢？互联网创业者大多有些赌徒的精神，那么这种赌徒精神算不算是互联网企业的原罪呢？

其实深究下去，互联网行业欺骗消费者、绑架用户、窃取隐私等一系列没有法律规定的逐利行为，这些东西堆积在一起无不是互联网原罪，终究有一天互联网行业会迎来它的清洗和调整，谁又能料到呢？只是希望处于浪潮之巅的我们认识到约束我们的除了法律，还有心中的道德标准和头顶的朗朗晴空。

回顾过去的几年，从贵阳大数据交易中心成立以来，大量的数据公司完成了从无到有的过程。2014 年做征信找数据除了身份验证和学籍验证的数据接口很难找到其他数据，数据交易许多时候仍旧停留在 Excel 表格线下传输的方式。反观现在市场上的数据公司，从人脸识别到设备指纹，从身份验证到银行卡交易流水，从手机号通话详单到历史浏览与阅读等，这么多的数据呈现爆炸式的增长，不仅如此，现在的数据都是以 API 接口的形式进行交易和传输，免费的数据交换普遍存在于这个行业，让人不禁感慨世界变化得真快呀！

随着数据公司爆炸式的增长，数据分析师群体增长的速度着实缓慢，目前市场上还没有较为成熟的数据分析师协会或者相关机构，大数据时代的领军者们还在学习的道路上，一起加油！

# 第10章 数据分析测试题与答案

## 10.1 MySQL 测试题

（1）试想数据库中有如下几张表格。

学生表：学号、姓名、性别、年龄

课程表：课程编号、课程名称

选课表：学号、课程编号、成绩

要求使用 MySQL 查询学生表中某学生选课的课程名称及成绩。

（2）试想数据库中有如下几张表格。

管理层表：部门、部门负责人

全体人员表：姓名、部门

要求使用 MySQL 列出每个基础员工的上司。

（3）试想数据库中有如下表格

交易表：交易日期、订单号、订单金额

要求使用 MySQL 查询 2015 年每月的最大交易量和最大交易额。

**答案：**

（1）

```
select a.姓名,c.课程名称,b.成绩
from 学生表 a
left join 选课表 b on a.学号=b.学号
left join 课程表 c on b.课程编号=c.课程编号
where a.姓名='张三'
```

（2）

```
select a.姓名,b.部门负责人 as 上司
from 全体人员表 a
```

```
    left join 管理层表 b on a.部门=b.部门
    where a.姓名 not in
    (select 姓名 from 管理层表)
```

（3）

```
    select month(日期) 月份,max(单量) 最大交易单量,max(交易额)
最大交易额
    from(
    select date(交易日期) 日期,count(订单号) 单量,sum(订单金额)
交易额
    from 交易表
    where 交易日期 like '2015%'
    group by 日期) a
    group by 月份
```

## 10.2 逻辑题

（1）一根棍子被折成了三段，三段能拼成一个三角形的概率是多大？

（2）赌场用一副常规扑克（52 张）玩一个游戏。规则是每次抽两张牌，如果两张都是黑色归庄家，如果两张都是红色归玩家，如果是一张红色一张黑色就被丢弃。这个过程持续直到 52 张牌抽完。如果玩家比庄家的牌多则赢 100 元，反之（包括平局）玩家什么都得不到。你愿意付多少钱玩一次这个游戏？

（3）两根蜡烛每根都能烧一个小时，但是由于密度不均他们烧得时间有时快有时慢，你怎样用这样的两根蜡烛计算 45 分钟？

（4）假设有一个矩阵数据有三列数据如表 10-1 所示。

表 10-1

| 城　　市 | 商　　家 | 销售额 |
|---|---|---|
| 1 | A | 1000000 |
| 1 | B | 1000000 |

续表

| 城　　市 | 商　　家 | 销售额 |
|---|---|---|
| 2 | C | 800000 |
| 2 | D | 1200000 |
| 2 | E | 900000 |
| … | … | … |

希望你用 SQL、R、Python、Matlab 等任意熟悉的语言生成一个新的表格，记录每个城市中销售额最大的商家名和最大的销售额。

（5）一副扑克牌（52 张），解释一下为什么抽到顺子（连续五张数字相连，从 2~6 到 10~A）的概率大于三加二（三张数字一样加两张数字一样，比如 AAA+33）。

**答案：**

（1）设长 L 的棍子任意折成 3 段的长度分别是 $x$,$y$ 和 $z=L-(x+y)$，其中：

$0<x<L$，$0<y<L$，$0<L-(x+y)<L$，即 $y<L-x$，

覆盖了图 10-1 的区域 1、区域 2、区域 3、区域 4 这四个部分的总体。

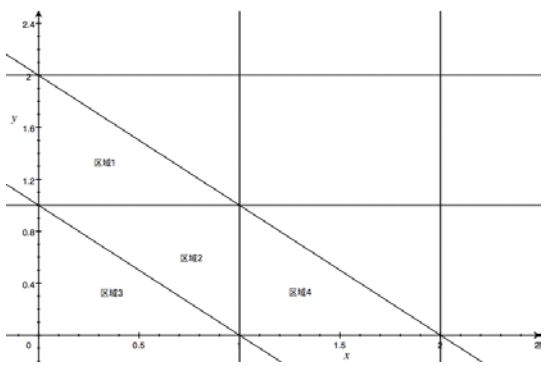



图 10-1

要求三段能构成三角形，则：

$x+y>z$，即 $x+y>(L-x-y)$，$y>L/2-x$

$y+z>x$，即 $y+(L-x-y)>x$，$x<L/2$

$z+x>y$，即 $(L-x-y)+x>y$，$y<L/2$

这三条公式覆盖了图 10-1 中区域 3 的位置，所求概率等于 $x+y=L/2$、$x=L/2$、$y=L/2$ 三条直线所包围图形的面积除以直线 $(x+y)=L$ 与 $x$ 轴、$y$ 轴所包围图形的面积。

故长 $L$ 的棍子任意折成 3 段，此 3 段能构成一个三角形的概率是：

$(L/2\times L/2\times 1/2)/(L\times L\times 1/2)=L\text{^}2/8/(L\text{^}2/2)=1/4$

（2）我们用 P11 代表第一次全黑概率，P12 代表第一次全红概率，P13 代表第一次一黑一白概率，P21 代表第二次全黑概率……

第一次都是黑色的概率：$P11=1/2\times 25/52$

第一次都是红色的概率：$P12=1/2\times 25/52$

第一次一红一黑的概率：$P13=1\times 25/52=2\times P11=2\times P12$

第二次抽取：

$P(21|11)=P(22|12)$

$P(21|12)=P(22|11)$

$P(21|13)=P(22|13)$

……

$E(P_{n1})=E(P_{n2})=1/2\times E(P_{n3})$

所以你有 1/4 概率赢，因此你最多为此游戏付出 100×1/4=25（元）

（3）点燃蜡烛 1 两端，同时点燃蜡烛 2 一段，蜡烛 1 燃完需 30 分钟，这时蜡烛 2 恰好燃了一半，再点燃蜡烛 2 的另一端，燃完是 15 分钟，加起来 45 分钟。

（4）

```
select 城市,商家,销售额
from(
select 城市,商家,销售额
from table
order by 销售额 desc
) a
group by 城市
```

（5）抽到顺子的概率：

总共有 52 张牌，因此抽取 5 张牌的组合数有 C(52，5)。既然要成为顺子，那么顺子必须为 2~6，3~7，…，10~A 这 9 种顺子，但是对于每张牌都有 4 个花色，故成为顺子的个数总共有 9×4^5 种顺子。因此抽到顺子的概率

$$P=9\times4^{\wedge}5/C(52,5)=0.35\%$$

抽到三加二的概率：

同样的 52 张牌抽取五张的组合数总共有 C(52，5)。要成为三加二其中“三”的种类有 C(13，1)个，四个花色随机取三个 C（4,3），“二”的种类有 C（12，1），四个花色随机去两个 C（4，2）。因此抽到三加二的概率

$$P=C(13,1)\times C(4,3)\times C(12,1)\times C(4,2)/C(52,5)=0.14\%$$

因此抽到顺子的概率大于三加二的概率。