
第 1 章
底层认知

# 1.1 基础认知

## 第 1 问：数据分析怎么学？——本书学习指南

### 1. 如何阅读本书?

对于从 0 到 1 的入门学习，重要的是要先建立对该领域的认知框架，再逐步进行知识的汲取。对于数据分析初学者而言也同样如此，这样的做法有如在学生时代学习时，老师根据教学大纲给学生授课，学生基于考试大纲进行复习以应对考试。

本书说不上是数据分析领域的权威大纲，但内容是基于笔者团队多年实战经验沉淀的知识框架。入门数据分析师可以围绕本书，针对不同方面的内容进行学习与实践。

在第 55 问“数据分析没有思路怎么办？——数据分析中‘以终为始’的思考逻辑”中，我们介绍了**以终为始**的思考逻辑，而它也同样适用在阅读中。这个逻辑建议读者带点“功利主义”去阅读，以便更高效地汲取书中知识，达成自己的目的。所谓的“功利主义”实际上也就是在做目标管理：这本书是在为你的哪些目标做服务？定义好阅读目标后，第二步才开始制订阅读计划：本书的内容如何帮助你实现目标？

而在此之前，需要先对本书的整体有个认识，也就是要做“检视阅读”。

**（1）检视阅读：了解一本书的大致轮廓。**

检视阅读建议读者先对所有内容进行快速略读，略读的目的是了解这本书的大致轮廓。因此，还没接触过数据分析的读者，对于一些不太能理解或者读不懂的地方，可以先跳过。

另外，目录是书的骨架，本书每个章节的前后逻辑都经过了笔者团队的反复讨论与优化。读者可以结合前言提到的胜任力模型来浏览目录，对本书内容有总体的认知。

了解了本书的大体内容后，可以按目的进行深入的分析阅读与主题阅读。

**（2）分析阅读：反复咀嚼、理解一本书，把书中知识变成自己的。**

这个阶段需要根据阅读目标，选择需要深入阅读的部分进行重点学习。在阅读过程中遇到不懂的地方，除了进行反复咀嚼外，还可以通过外部知识补充理解。

一本书的价值是作者与读者之间的相互成就。为了达到“把书中知识变成自己的”这个目的，需要读者通过“输出”来倒逼知识的“输入”（费曼学习法）。

读者按目的选择需要深入阅读的部分后，可以输出笔记：

- 结构笔记：全书围绕着初级数据分析师胜任力模型展开，为了能让读者理解这个

模型，建议读者能主动输出结构笔记，笔记的重点是数据分析的能力结构，而不是细节知识。

- 概念笔记：在业务实践或者回答业务面试题的过程中，形成自己的分析框架；分析框架的搭建可以参考下文“搭积木”的方法实现。

**（3）主题阅读：在一个主题下做延伸阅读。**

相比于分析阅读，主题阅读处在更高的阅读层次。什么叫主题阅读？顾名思义，就是带着一个“主题”，或者说带着“解决某个业务问题”的目的来看本书。

在前面阅读步骤的沉淀下，如果已经对数据分析中的大部分知识点有了一定了解，此时为了达成主题阅读的目的，可以借助“搭积木”的方法使用本书。

### 2. 如何使用本书？——“搭积木”的方法玩转本书

什么是“搭积木”？在理论学习阶段，需要读者通读本书，对数据分析建立全面的认知，搭建起自己的分析“武器库”。在实践阶段，遇到问题，采用类似“搭积木”的方式，在“武器库”中选择合适的“武器”（思维方法、分析工具）解决问题。

对于初学者而言，在分析实践过程中，有个常见的问题：该如何选择分析方法？基于笔者团队的业务经验，有一个重要的技巧，就是先回答“业务需求方是谁？”这个问题，由此基于不同部门得出不同的分析方法：

- 用户运营部门、会员管理部门：从常见的用户分析方法入手（第 32 ～ 35 问）。
- 产品部门、产品经理：从常见的产品分析方法入手（第 29 ～ 31 问）。
- 市场部门、战略部门、产品部门：从常见的行业分析方法入手（第 23 ～ 28 问）。

以上就是简单的“搭积木”方式：根据实际的业务场景，从书中挑选出合适的方法论丰富“武器库”。更进一步，深入数据分析万能流程中，可以按如下方式“搭积木”：

**（1）明确问题。**

这个阶段重要的是对业务问题进行清晰定义。借助积木可完成以下工作：

- 数据思维的逻辑整理（第 12 ～ 14 问）；
- 描述性分析（第 16 问）；
- 对比分析（第 17 问）。

**（2）分析原因。**

这个阶段重要的是对前面定义好的业务问题进行下钻分析。借助积木可完成以下工作：

- 数据异常分析（第 15 问）；
- 归因分析（第 19 问）；
- 预测分析（第 20 问）；
- 相关性分析（第 21 问）。

**（3）落地建议。**

这个阶段重要的是能给出落地（即业务可操作）的有效建议。借助积木可完成以下工作：

- 了解业务（第 58 ～ 63 问）；
- 给出落地建议（第 67 问）。

这些积木能组成一个完整的分析流程“武器库”，当然这也只是一种方案。前面阅读本书的建议中提到分析阅读时，做概念笔记就是要形成自己的分析框架。而选择积木的过程，就是读者形成自己分析框架的过程。由此可见，**搭积木没有标准答案，会产生不同的组合方案。**而正是这些方案能适应不同的业务场景需求，对业务问题进行分析、解决。

读者从目录可以看到本书的内容非常丰富，但想进入一门学科，或者说想胜任一个岗位所需的知识却远不止这些。受限于篇幅，**笔者团队准备了一份与本书搭配使用的小册子，请扫码获取。**

读者还可以通过关注微信公众号木木自由、数据分析星球、饼干哥哥数据分析，回复“72 问小册子”获取。此外，本书勘误、知识加餐等内容也会放在小册子中。

小册子主要为实战服务，里面有许多开箱即用的代码。对于初学者而言，最好的学习途径就是“先模仿，再创作。”因此，读者可以阅读小册子，并按照教程进行代码工具的安装。然后把本书及小册子涉及的代码都跟着敲一遍，最后你会很神奇地发现对代码没有了抗拒，并且在持续的学习过程中会逐渐熟悉它们，甚至能用它们来提高工作效率。

## 第 2 问：数据分析是怎么来的？——数据分析极简发展史

> 导读：为了深刻认识数据分析，有必要对它的来龙去脉进行一番讨论。讨论来龙去脉不是为了考察数据分析的国内外发展史，而是从数据分析的发展中探索本质，建立底层认知。

### 1. 了解数据分析发展

从游牧时代开始，就已经涉及数据分析了。例如，今天抓了一只野猪，明天抓了一只羊，所以猎物总共有两只，如何分配呢？羊可以养起来，因为羊可以产奶，给孩子补充营养；猪可以杀掉，一天吃不完，那就分两天吃，首领多分一些，其他人少分一些……这正是数据分析的早期应用。可见，数据分析的历史很悠久，可以说在人们开始使用数字的时候就已经有数据分析的意识了。

在过去的十年到二十年里，数据分析一直是非常热门的词汇，但是在更早的生产活

动中，数据分析其实就已经存在了，只是那时主流市场并未产生需求。那数据分析是怎么成为咨询公司麦肯锡所说的“重要的生产因素”的呢？换句话说，热门的数据分析岗位是怎么产生的呢？

从下图的阿里发展史中，我们可以看到这样的发展路径：

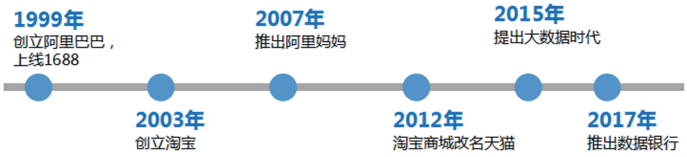
（1）阿里创立自己的产品——1688 网站；

（2）初创团队的成员开始联系批发贸易商入驻，即开展销售业务及网站运营工作；

（3）随着业务的发展，为满足市场需求，除了对现有产品进行迭代优化外，阿里还推出许多的产品：淘宝、天猫等，这背后需要有专业的产品经理支持，提高业务运营流程效率；

（4）随着规模的扩大、数据的积累，专业数据分析师的需求应运而生，借助数据分析、数据挖掘的方法论优化产品迭代、业务增长策略；

（5）随着数据的使用场景日趋成熟，数据使用需求也越来越大，需要通过衍生的数据产品来优化数据分析流程效率，如数据银行、达摩盘、策略中心。




当然这不是严谨的发展史，例如数据挖掘技术早在 20 世纪 90 年代就存在了，这里的发展路径更多是从主流市场的角度来理解，也可以说是求职市场的变化。例如 2015 年以前很少有专门的数据分析师岗位，后来随着大数据在工业界的普及、落地，市场对数据分析师的需求多了起来。再例如数据产品经理也是随市场的发展而兴起的。

### 2. 窥探发展路径背后的业务场景需求

从数据分析的发展路径中，我们可以进一步去窥探其背后业务场景需求的变化：

在发展初期，市场还处在“开荒阶段”，那时的产品比较简单，对应的运营玩法也比较简单，此时体系不完善，主要**依赖经验、直觉来驱动业务增长**，例如之前没有做广告投放，现在做了，效果就有了。

在发展中期，为了追求规模化，品牌需要不断去扩展边界，于是基于现有运营能力，把成功经验复制到其他细分市场的模式就很重要，进而成体系的运营方法论、产品方法论需求应运而生，也就是要**从以往经验中沉淀出泛化能力强的业务模型框架，来实现增长**。例如以往做用户运营，尝试过用近期消费距离、累计消费频次、累计消费金额来做用户分层运营，效果不错，因此可以把方法论总结成 RFM 模型应用到更多场景中。

度过了“野蛮生长”的增量时代后，市场竞争格局形成，竞争对手运营体系成熟，

再想从增量市场抢夺用户成本将变得很高，而且手里的存量客户如果没有及时维护也容易被竞争对手夺去；**此时的业务需要更精准的方法来指导决策，于是代表理性、客观的数据登上舞台，数据分析就变得很重要**。例如运营中常说的“魔法数字”：利用数据分析方法计算 RFM 模型的特征阈值，能够得到更精准、有效的分层模型。

### 3. 小结

从数据分析的来源中我们可以看到，**数据分析的定位从来都不是“雪中送炭”，而是在发展到一定程度，有了夯实基础之后的“锦上添花”**。此外，对数据分析来源的讨论，是为了说明一件事：数据分析并不独立，它来源于业务，最终又在业务落地。所以**想做好数据分析一定要懂业务**，否则不论是分析逻辑还是最后的赋能建议都无法落地，无法实现数据分析价值。

## 第 3 问：什么是数据指标?

> 导读：了解完数据分析的发展后，本问开始，将从数据分析的核心——“数据指标”切入，建立全面的数据分析底层认知。数据指标是业务现状的反映，而数据分析也正是基于对业务现状的准确透视才能做出有效决策，因此，数据指标的重要性不言而喻。

为了建立对数据指标的完整认知，我们把数据指标拆成“数据”与“指标”，指标是数据之间的运算，是“衡量”事物发展程度的“模型”。也就是说通过“建立指标”评估“业务发展”是一个建模的过程，是把业务发展从物理世界映射到数据空间，只有这样才能使得“万物皆可计算”，这就是数据分析的基础。

为了厘清从数据到指标的建模过程，我们需要先对“数据”的概念进行讨论。

### 1. 什么是“数据”？

数据是被存储起来的信息。从应用的角度看，数据是把事物做量化处理的工具。万物皆可数据化，数值是数据，文本、图像、视频等同样也是数据。

（1）按字段类型划分，可以把数据分为：

文本类：常见于描述性字段，如姓名、地址、备注等。

数值类：最为常见，用于描述量化属性，如成交金额、商品数量等。

时间类：仅用于描述事件发生的时间，是重要的分析维度（如同比、环比、累计等）。

（2）按结构划分，可以把数据分为：

结构化数据：通常指以关系数据库方式记录的数据。

半结构化数据：如日志、网页数据。

非结构化数据：如语音、图片、视频等形式的数据。

（3）根据数据连续的属性不同，可以把数据分为：

连续型数据：在任意区间可以无限取值，例如年龄、身高。

离散型数据：常见于分类数据，例如性别、年级。

## 2. 如何理解“指标”？

指标的作用是“度量”业务，可以从三个角度对指标进行拆解：指标 = 维度 + 汇总方式 + 量度。

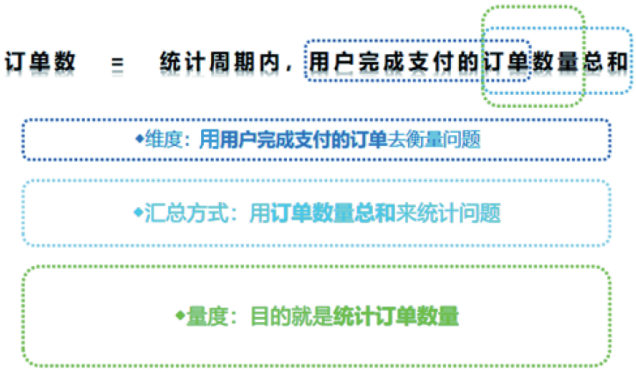
维度：从什么角度去衡量问题。

汇总方式：用什么方法去统计问题。

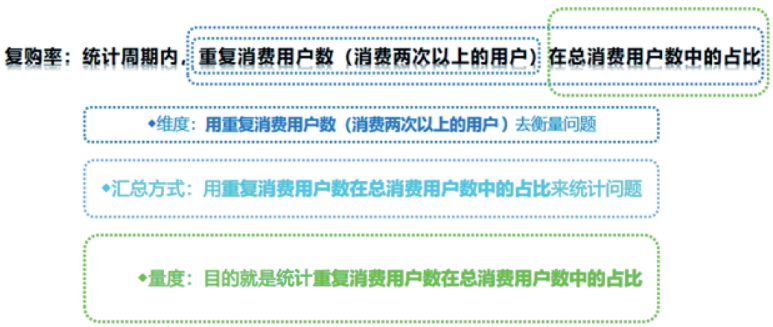
量度：目标是什么。

下面举两个例子。

订单数是指统计周期内，用户完成支付的订单数量总和。从维度 、汇总方式、量度三个角度将订单数拆解，如下图所示。




复购率是指统计周期内，重复消费用户数（消费两次以上的用户）在总消费用户数中的占比。从维度 、汇总方式、量度三个角度将复购率拆解，如下图所示。



### 3. 数据指标如何落地使用?

了解完指标的底层逻辑（理论）后，更重要的是将指标在业务中落地。笔者团队结合数据分析经验，总结了以下数据指标的落地建议：

**（1）指标基建——确保数据的完整、准确。**

为了打下指标模型的稳固基础，需要对数据底层进行检视：

① 检视数据源头：埋点收集的事件数据是否足以支撑所需指标的建模。

② 脏数据清洗逻辑，也就是数据仓库中常见的 ETL（Extract-Transform-Load，抽取 - 转换 - 加载）概念。

**（2）从业务层面理解指标。**

理解业务是数据分析落地的前提，有效地使用指标也同样如此，要求熟悉数据指标背后的业务含义。例如“会员成单数”这个指标本身有很多含义（针对以购买会员为主要商业模式的 App）：

- 直接含义：整个团队的业务完成能力。
- 会员成单数 + 成本：企业的盈利能力。
- 会员成单数 + 产品：产品畅销程度。
- 会员成单数 + 用户分层：用户的需求。

**（3）从指标的变动中做决策。**

为了判断业务现状的好坏、趋势，需要建立衡量标准，数据指标的使用同样如此。

通过某个孤立的指标不能反映现实，例如小明身高 165cm，我们看不出小明的身高特征，但是当走来一个身高 180cm 的人时，我们就能判断小明相对比较矮，或者当我们拿到全国平均身高水平是 167cm 时，也能得出同样的结论。这就是利用对比思维建立标准的过程，对比的客体可以是横向的同属性对象、总体平均，也可以是纵向的历史数据。

如果是周期性变化，那很有可能是正常波动，可以初步判作“正常”。如果是“突发 + 下跌”，那很有可能是异常的波动，可以初步判作“问题”。

**（4）指标的生命周期——不同阶段使用不同指标。**

既然指标的作用在于反映业务，而业务的发展存在生命周期，那指标的使用也应存在时效性，即指标的生命周期。

沿着产品的生命周期来看，不同阶段使用的指标差异如下：

① 导入期：业务目标在于建立知名度，通过口碑引流，着重关注新注册人数、分享率指标。

② 成长期：业务目标在于通过不同渠道布局推广最大限度占有市场，着重关注新会员来源渠道占比等指标。

③ 成熟期：业务目标在于将前期流量变现，确保盈利规模，着重关注付费率、毛利率等指标。

④ 衰退期：此时，市场增量收缩，要求对存量人群精细化运营，着重关注复购率、重购金额占比等指标。

### 4. 小结

在一定程度上，“数据指标”能揭示出产品用户的行为和业务水平状况。当然，我们也不能完全迷失在数据中，应注意以下几点：

- 数据不等同于实际场景，实际场景往往比数据更加复杂，分析时需要了解具象化的场景，而不是抽象的数据。
- 数据本身没有观点，分析时不能预设观点，只倾向于那些能够支持自己观点的数据。
- 数据具备一定的时效性，不同情况下，一些曾经的数据可能不再适用，需要找到新的数据指标。

总之，精确的数据无法代替大方向上的判断，不要过分迷恋数据，要做到具体问题具体分析，形成发现问题、分析问题、总结问题、解决问题的思路闭环。

## 第 4 问：常见的指标有哪些？

> 导读：为了帮助读者对数据指标有更直观的认识，本问将介绍常见行业的指标体系。前面我们说指标可以反映业务现状，但“隔行如隔山”，不同领域的业务之间存在明显差异。了解目标领域常用的指标，可以帮助我们快速熟悉业务。

### 1. 互联网行业

互联网产品具有边际成本低、传播速度快等特点，由此造就了互联网产品用户量大、使用频率高、迭代速度快等优势。这样的业务场景下，数据分析能有更多的落地场景，因此经典书籍《增长黑客》里的增长方法论、案例等都是基于互联网产品展开的。

这里的互联网产品主要指的是 C 端的 App、网站甚至是游戏（本质也是 App）等，虽然不同行业的产品服务的人群、场景不同，例如滴滴服务的是出行场景，而淘宝服务的是购买场景，**但它们的底层逻辑是相通的，也就是可以借用同一套指标体系来进行数据分析。**只是在具体落地应用时，不同的场景会关注不同的数据指标。**请在本书前言扫码获取小册子，查看互联网行业常见的指标及定义。**

### 2. 零售行业

与互联网相比，零售行业显得更传统一些，但是在数据使用场景上，以沃尔玛为代表的大型零售商高度依赖数据对其供应链、选品等方面进行赋能提效。以淘宝为代表的

电商行业，从 1999 年发展至今，已经积累了庞大的数据量，并在电商流程上形成了成熟的数据解决方案，帮助商家提高销售额、优化买家用户体验。

大数据时代产生了“人、货、场”的新零售概念，笔者团队则按该逻辑，为读者展示零售行业的数据指标体系全貌。零售行业常见的指标及定义详见小册子。

### 3. 金融行业

与互联网、零售行业相比，金融行业的平稳运行特别依赖大数据，因此，找到更有效的数据指标以及分析方法非常重要。“数据分析”也为金融行业重塑业务提供了更多的、更广泛的思路和策略。

例如，金融部门进行风险管理、欺诈检测，识别数据中的异常或不良模式，并指示公司的安全部门采取适当措施降低风险。从金融消费者行为实时分析中获得有价值的见解，有助于改善个性化服务，以增加销售额并衡量客户的生命周期价值等。财务方面，则需要更加积极地运用“数据分析”来保护客户利益并促进金融服务行业的发展。金融行业常见的指标及定义详见小册子。

### 4. 小结

做数据分析会遇见很多指标，我们应该清楚哪些要着重分析，哪些指标最契合当下的分析需求。注意，具体到不同业务，不同指标的定义可能略有差别，但是思路是一致的。

## 第 5 问：对于数据分析领域，统计学要学到什么程度？

> 导读：翻开贾俊平老师的《统计学》教材，400 页的信息扑面而来，内容包括图形信息化、数据的集中趋势、概率计算、排列组合、连续型概率分布、离散型概率分布、假设检验、相关和回归等诸多复杂的知识点。初学者时常大呼“难学”，但实际上，学习是有“捷径”的，那就是“以终为始”——根据目标场景需求制订学习计划。那么，对于数据分析领域，统计学要学到什么程度呢？

### 1. 什么是统计学?

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型科学。而数据分析是基于统计方法研究数据，其所用的方法分为描述统计和推断统计。

（1）描述统计。

描述统计是研究一组数据的组织、整理和描述的统计学分支，内容包括取得研究所

需要的数据，用图表形式对数据进行加工处理和显示，进而通过综合、概括与分析，得出反映所研究现象的一般性特征。

描述统计主要应用在探索性数据分析阶段（Explore Data Analysis，EDA），在分析之前先对数据的结构、分布等特征进行了解，从而制订数据清洗、特征工程等方案。

**（2）推断统计。**

推断统计是研究如何利用样本数据对总体的数量特征进行推断的统计学分支，其内容包括抽样分布理论、参数估计、假设检验、方差分析、回归分析、时间序列分析等。

描述统计最经典的应用场景就是 AB 测试、销售预测。

### 2. 如何开始?

开始学习统计学最重要的是从宏观上有一个初步的认识，如统计学大概包括哪些内容、能够做什么、解决哪些问题等，然后再深入细致地去了解它，这样的话，你在学习每一部分知识时，就能够清楚地知道该部分知识的地位和作用。接着以“搭积木”的思维，从基础开始，层层递进。最后在深入学习的时候，一定要结合自己目前的需求，有所侧重。

**（1）推荐教材。**

统计学相关的推荐阅读教材如下所示。

| 书名 | 作者 | 特点 | 使用场景 |
| --- | --- | --- | --- |
| 《深入浅出统计学》 | 作者：道恩·格里菲思<br>译者：李芳 | 结合图像和小例子的形式进行讲解，阅读轻松 | 入门 |
| 《赤裸裸的统计学》 | 作者：查尔斯·韦兰<br>译者：曹槟 | 这本书有生动诙谐的案例，通俗易懂，图文并茂，学习统计学不会那么枯燥 | 入门 |
| 《统计学：从数据到结论》 | 吴喜之 | 没有复杂的公式，不过内容讲得很通透。内容不死板，一本小书一天就看完 | 入门 |
| 《大话统计学》 | 陈文贤、陈静枝 | 本书前后连贯，各章之间也是先后呼应。可以从零开始接触统计学，并将其真正应用到工作中 | 入门 |
| 《应用统计学》 | 张梅琳 | 从实用场景出发的高频统计学知识点，3 ～ 4 个小时就能看完 | 进阶 |
| 《统计学》 | 贾俊平 | 数学原理讲解完整 | 深入 |
| 《统计学习方法》 | 李航 | 与机器学习结合 | 深入 |

**（2）针对数据分析，统计学要学到什么程度？**

**从广度来看：**

首先要了解一些统计学的基本概念，例如描述型统计、假设检验、正态分布，然后再去学习统计学里的数据模型，例如聚类、回归，这些都是业务分析中必备的内容。

大部分的数据分析，都会用到以下统计学的知识，可以重点学习，而且这一部分概念简单，很容易掌握：

- 基本的统计量：均值、中位数、众数、方差、标准差、百分位数等。
- 概率分布：几何分布、二项分布、泊松分布、正态分布等。
- 总体和样本：了解基本概念，如抽样的概念。
- 置信区间与假设检验：学会如何进行验证分析。
- 相关性与回归分析：一般数据分析的基本模型。
- 数据展示图形（8 种基础图形）。

以经典教材《统计学》为例，笔者团队对内容按入门、进阶进行了划分，对大多数初学者而言，仅需学习入门内容即可。随着数据分析工作的深入，对分析能力有拔高要求的读者，可以进一步学习进阶内容。**请在本书前言扫码获取小册子，查看统计学入门与进阶目录。**

**从深度来看：**

前面说过知识点的学习需要“以终为始”，从需求场景出发，有落地应用场景的知识点才有必要深入学习，否则即使学习了，无用武之地也很容易忘记。对于初学者而言，重要的是掌握统计学的概念，不需要深究原理，但要知道如何“查看”及“应用”统计结果。

那只知道概念，不知道原理的话，在工作中要如何实践呢？实际上，绝大部分统计学的知识已经被封装成了开箱即用的工具。也就是说，相比于数学原理，实践中更重要的是会使用工具。例如使用 Excel 时，能利用它实现相关性分析、回归分析等复杂方法即可。对于进阶的工作内容，可能更多使用 Python 工具。同样，学会调包、调参即可满足 90% 的应用场景。

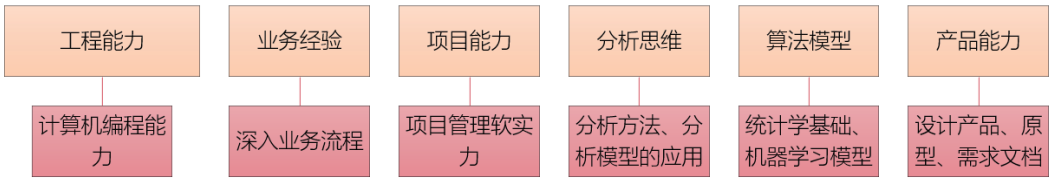
但是有一个场景是例外，那就是面试。我们常说“面试造火箭，工作拧螺丝”，尽管实践中能解决问题即可，但面试仍会要求我们懂得统计学高频知识点背后的数学原理。

### 3. 小结

统计学是一门交叉性和应用性都很强的学科。统计学源于实践并用于实践，通常从实际应用问题开始，经过加工提炼，形成概率统计模型，并最终指导实践。一个问题的完整解决往往需要设计试验、数据处理分析、撰写总结报告等。因此，统计学是一名优秀数据分析人员必须具备的知识。

## 第 6 问：数据分析领域主要的岗位有哪些？

导读：随着大数据的兴起，数据分析相关的招聘也越来越多，但很多人对该领域的很多职位和工作内容仍然不是很了解。目前，数据分析领域主要有以下几类岗位：业务数据分析师、商业数据分析师、数据运营、数据产品经理、数据工程师、数据科学家等，按照工作侧重点不同，本问将上述岗位分为偏业务和偏技术两大类，并对每个岗位按照下图所示技能栈进行分析，阐述不同岗位的特点。



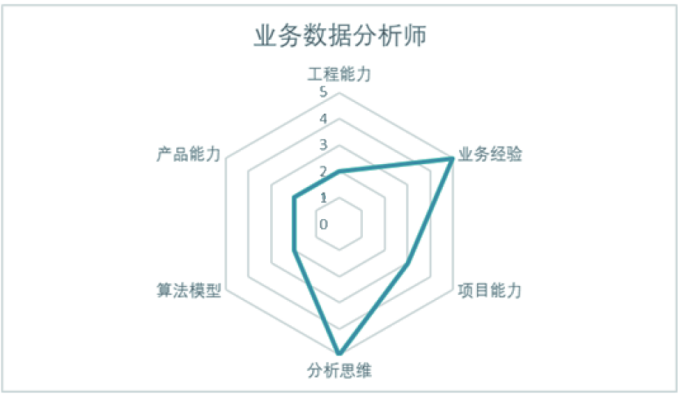

### 1. 偏业务方向的数据分析岗位

偏业务方向的数据分析岗位一般归属于业务部门，有业务数据分析师、商业分析师、数据运营、数据产品经理等，该类岗位的职位描述如下图所示。


**（1）业务数据分析师。**

业务数据分析师需要将业务数据体系化，建立一套完善的指标体系，并完成数据提取、清洗、多维度分析及预测等工作，并生成策略推动落地。数据分析师可以基于指标体系进行拆解，逐层细化，抽丝剥茧，找到问题的根因。指标体系如果需要自动化监控，还需要进行 BI 报表开发，所以数据分析师也需要了解一些 BI 工程师的知识。

该岗位所要具备的技能栈如下图所示。

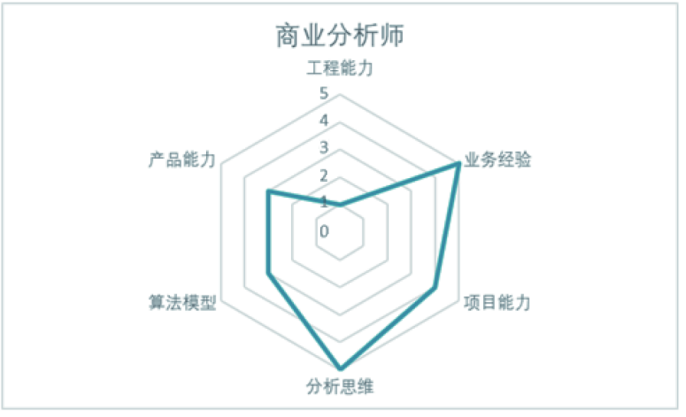



**（2）商业分析师。**

商业分析偏向经营和战略方向的分析，一般更加宏观，通常涉及业绩目标制定、各个渠道经营状况监控、业绩指标异常监控和量化归因并为决策者提供决策依据，同时还需要有敏锐的商业嗅觉，对市场、竞对有较为全面的认知，能快速察觉政策、竞对、市场风向等，并及时做出响应。

例如，想要开一家快递驿站，首先需要考虑在哪里开，这就要调查居民密度、居民消费能力、竞争对手、线上消费能力等因素。这些分析更加宏观，数据来源广泛，而且需要一些调研进行定性研究，和业务数据分析这种微观的分析有一些差异。

该岗位所要具备的技能栈如下图所示。

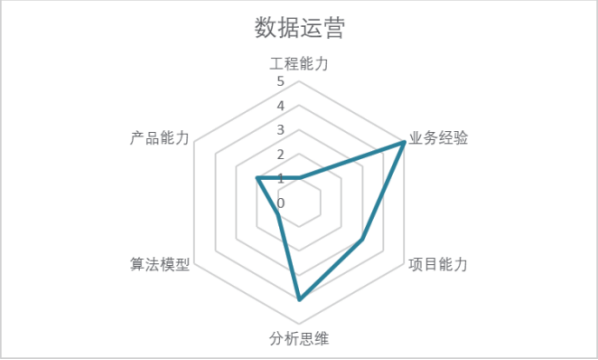



**（3）数据运营。**

数据运营主要负责运营相关数据的分析，为日常运营提供数据支持，协助运营人员制定运营策略和方案落地。

以活跃指标的下跌为例，需要分析的问题有：活跃指标下跌了多少？是属于合理的数据波动，还是异常波动？什么时候开始下跌？是整体的活跃用户下跌，还是部分用户？为什么下跌？是产品版本迭代，还是运营效果不佳？怎么解决下跌的问题？

该岗位所要具备的技能栈如下图所示。



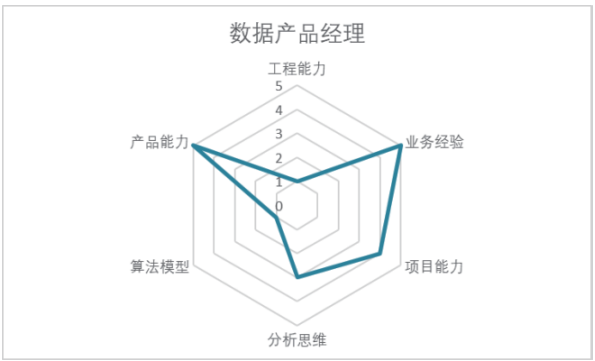

（4）数据产品经理。

这个岗位比较新，要求同时具备产品经理和数据分析师的技能。它有两种定位：一种是具备强数据分析能力的产品经理，另一种是公司数据产品的规划者。

前者以数据为导向优化和改进产品。产品经理有更多的机会接触业务，可以顺便把数据分析师的活也干了，属于一专多能的典型。大到页面布局、路径规划，小到按钮的颜色和样式，数据产品经理都可以通过数据指标评估，擅长用分析进行决策。

后者是真正意义上的数据产品经理。随着数据量的与日俱增，会有不少与数据相关的产品项目，如大数据平台、埋点采集系统、数据可视化系统等。这些也是产品，但是更注重数据呈现，也需要提炼需求、设计、规划、项目排期，乃至落地。

该岗位所要具备的技能栈如下图所示。




## 2. 偏技术方向的数据分析岗位

偏技术方向的数据分析岗位有数据开发工程师、数据挖掘工程师、算法工程师等，该类岗位有的归属研发部门，有的则单独成立数据部门。与偏业务方向的数据分析岗位相比，偏技术方向的数据分析岗位要求有更高的数理知识以及开发能力。

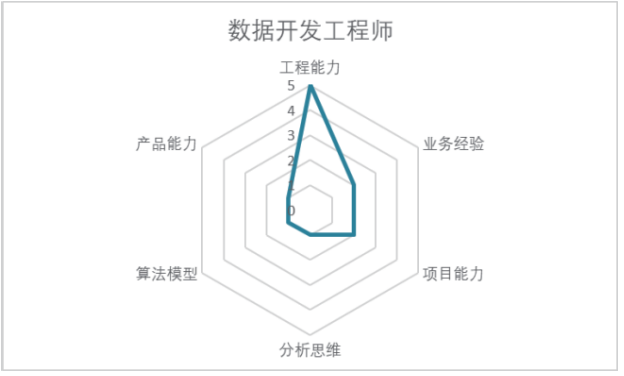
（1）数据开发工程师。

数据开发工程师更偏数据底层，其工作内容有数据采集、清洗、存储、建设数据仓库、数据应用、建设数据平台等。这个岗位基本不涉及数据分析的能力，而对大数据处

理能力要求较高，需要较强的编程及架构设计能力。

在很多中小型公司，由于人力有限，数据分析师还会承担一部分数据开发工程师的工作，兼做一部分数据清洗、ETL 和数据表开发的工作。

该岗位所要具备的技能栈如下图所示。

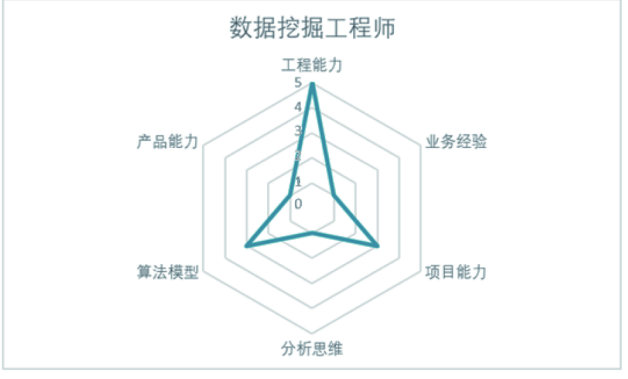



**（2）数据挖掘工程师。**

从概念上说，数据挖掘是通过一些数据挖掘算法（如分类、聚类、回归、预测、协同过滤、关联规则等）挖掘海量数据背后的业务价值。

如寻找共享单车最大效率的投放策略就是数据挖掘的工作范畴。数据挖掘工程师除了需要掌握算法基本原理，还需要很强的编程能力，如 Python、Scala、Java，往往也要求具备 Hadoop/Spark 的工程实践经验。单看工作内容，数据挖掘对分析能力没有业务型数据分析那么高，但这不代表业务不重要，尤其在特征选取方面，对业务的理解很大程度会影响特征的选取，进而影响模型效果。

该岗位所要具备的技能栈如下图所示。

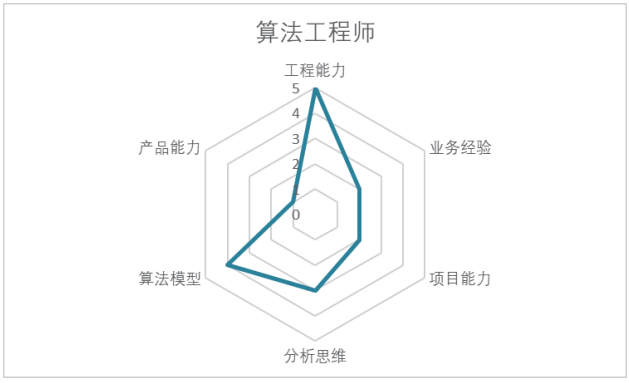



**（3）算法工程师。**

数据挖掘工程师可以继续精进成为算法工程师，后者对理论要求更高，不局限于简单的分类或者回归，还包括图像识别、自然语言处理、深度学习等领域。深度学习更前沿，它由神经网络发展而来。因为各类框架、模型较多，算法工程师除了要求熟悉

TensorFlow、Caffe、MXNet 等深度学习框架，对模型的应用和调参也是必备的，后者往往是普通“码农”和“大牛”的区别之处。

该岗位所要具备的技能栈如下图所示。




### 3. 小结

上面介绍了数据分析相关岗位的主要工作内容，以及不同岗位之间的区别，大家可以基于自己的兴趣和特长选择相应的岗位。一般来说，对于新人，比较适合的发展路线是先成为一名业务数据分析师，积累一定的经验后，再决定是向商业分析、数据挖掘方向发展，还是精进成为数据运营经理、数据分析经理等管理层。但无论是偏业务的岗位还是偏技术的岗位，要想借助数据驱动业务产生价值，必须是业务和技术并重，业务是终极目的，技术是实现业务的手段，两者相辅相成，缺一不可。

## 1.2 底层逻辑

### 第 7 问：如何建立完整有效的数据指标体系？

导读：清楚了什么是指标，了解了常见的指标，接下来就需要建立一个完整、有效的数据指标体系来帮助数据分析人员更好地梳理、理解业务，发现业务过程中出现的问题，进而推动业务的迭代优化。本问就从什么是指标体系、为什么需要指标体系、如何建立数据指标体系这三个方面介绍指标体系的基本概念和构建方法。

### 1. 什么是指标体系?

指标体系，即相互之间有逻辑联系的指标构成的整体，所以一个指标不能叫指标体

系，几个毫无关系的指标也不能叫指标体系。好的指标体系有如下特点：

（1）能体现当下业务的关注点。

我们知道，业务在不同时期的重心不同，关注的核心指标也会不同。例如，一般在业务初期会重点关注新用户的增长和留存；中期会关注用户的活跃和转化复购；后期会关注用户的流失和召回。不同阶段关注的核心指标不同，所对应的指标体系也必定有所差异，所以不能指望一套指标体系从头用到尾，每个阶段都应该针对当下的业务关注点搭建指标体系，这样才能够和业务保持一致，真正起到指标体系的价值。

（2）同时包含结果性指标和过程性指标。

我们习惯通过指标体系监控业务的发展趋势和出现的问题，但更重要的是，我们希望了解问题背后的原因，知其然更要知其所以然，对症下药才能够针对性地进行改进和优化。所以一个好的指标体系除了要有表征现状问题的结果性指标，还要有影响这个结果的过程性指标，这样才能在出现问题时有据可循，快速找到出现问题的原因。

（3）有对应的业务抓手。

在前面两点的基础上，我们除了希望通过指标体系反映业务现状、定位问题原因外，更希望它能够指导业务动作，告诉哪些部门应该在哪些环节进行改进，这个就是我们说的“要有对应的业务抓手”。要有具体到人、具体到策略的指导性意见，否则就算定位到了问题的原因所在，没有对应的人和策略跟进，问题依然得不到解决。

注意：建立指标体系不是一个人能够完成的，需要业务部门（市场、运营、产品部门等）、数据部门、开发部门相互协作，共同讨论确认。一个人闭门造车建立的指标体系很容易和业务脱节，也很难落地。在日常工作中，业务部门、数据部门、开发部门也需要紧密合作。

### 2. 为什么需要指标体系?

指标体系的作用如下：

（1）全面诊断业务现状。

没有指标对业务进行系统衡量，就无法明确业务现状，也就无法把控业务发展，尤其现在很多业务比较复杂，单一数据指标容易片面化。因此，搭建系统的指标体系，才能全面衡量业务发展情况，针对性地制定业务策略，促进业务良性增长。

（2）快速定位业务问题。

一个完整的指标体系能够明确结果型指标和过程型指标的关系，不仅能监控结果，更能分析过程。通过结果型指标回溯到和用户行为相关的过程型指标，找到解决问题的核心原因。如转化率这种结果型指标，影响它的可能是浏览次数、停留时长等过程型指标，通过指标体系，能明确转化率和浏览次数、停留时长的关系。

（3）有效驱动业务发展。

产品、运营、市场营销等部门都是促进公司发展的重要组成部分，而这些部门都需

要通过数据发现业务上的问题，针对性地提升改进。产品需要通过数据评估版本迭代效果；运营需要通过数据验证运营策略；市场营销需要通过数据洞察用户的消费习惯。通过完整的指标体系和数据分析，可以有效指导各部门的工作，通过数据找到业务当前痛点和瓶颈，以数据驱动找到优化方向，进而实现业绩的提升。

### 3. 如何建立数据指标体系?

一个指标体系的构建通常需要先确定一个核心指标作为一级指标，然后将核心指标进行逐层拆解，得到一个完整的指标体系。这里涉及几个关键的问题：如何确定这个核心指标？如何进行业务拆解？拆解后的过程如何进行衡量？这里介绍一种常用的构建指标体系的模型——OSM（Object Strategy Measure）模型，这个模型的含义如下：

O（Object，目标）：在建立数据指标体系之前，一定要清晰地了解当下的**业务重点和目标**，也就是模型中的 O。换句话说，业务的目标对应着业务的核心指标，了解业务的核心指标能够帮助我们快速厘清指标体系的方向。

S（Strategy，策略）：了解业务目标和核心指标之后，就需要在此基础上根据用户行为路径进行拆解，这个拆解一定对应着**业务策略**，也就是模型中的 S。把核心指标拆解成一个个过程指标，每个过程指标对应着相应的行动策略，这样就可以在整条链路中分析可以提升核心指标的点。

M（Measure，指标）：针对上面拆解的每个业务过程，制定对应的**评估指标**，也就是模型中的 M。评估指标的制定是将产品链路或者行为路径中的各个过程指标进行下钻细分，这里用到的方法就是麦肯锡著名的 MECE 模型，需保证每个细分指标是完全独立且相互穷尽的。

下面通过一个电商行业的案例来了解如何基于 OSM 模型构建指标体系。

**（1）明确核心指标。**

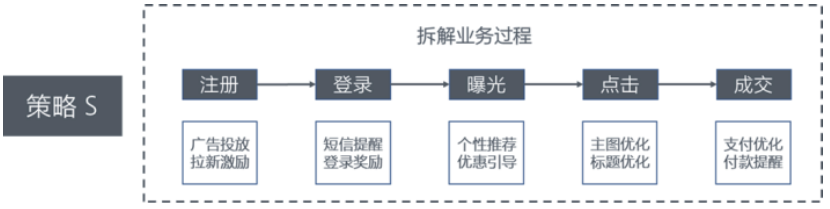
构建指标体系的第一步，需要明确当下业务的目标（Object）是什么，找到核心指标作为一级指标。例如当下的业务目标是增加营收，对应的核心指标就应该是总营收（Gross Merchandise Volume，GMV）。


**（2）拆解业务过程。**

明确了核心指标或者一级指标是 GMV，接下来就要对业务过程进行拆解，影响到 GMV 的各个环节有哪些？用户到最终付费贡献营收一般需要经历以下完整过程：注册产品→登录产品→商品曝光给用户→点击商品浏览详情→收藏加购→成交转化。

这样一来就把核心指标对应的中间过程梳理出来了，同时，针对每个中间过程也有对应的策略（Strategy），例如在注册环节，可以通过广告投放和优惠激励的形式进行拉

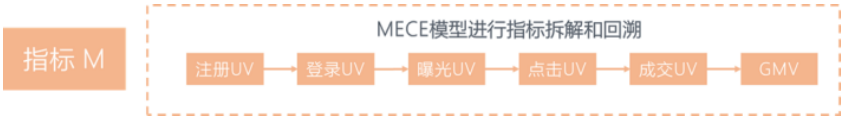
新、提高注册量等。



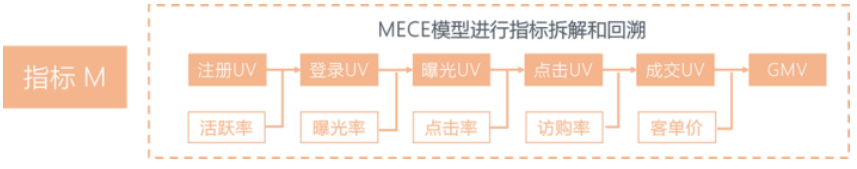

**（3）指标体系细分。**

对这些中间过程建立指标，并向下进行逐层拆解，这个过程我们称为指标体系分级治理，用到的模型是 MECE 模型。MECE 模型的指导思想是完全独立、相互穷尽，根据这个原则拆分可以逐层细化，暴露业务本质，帮助我们快速地定位业务问题。

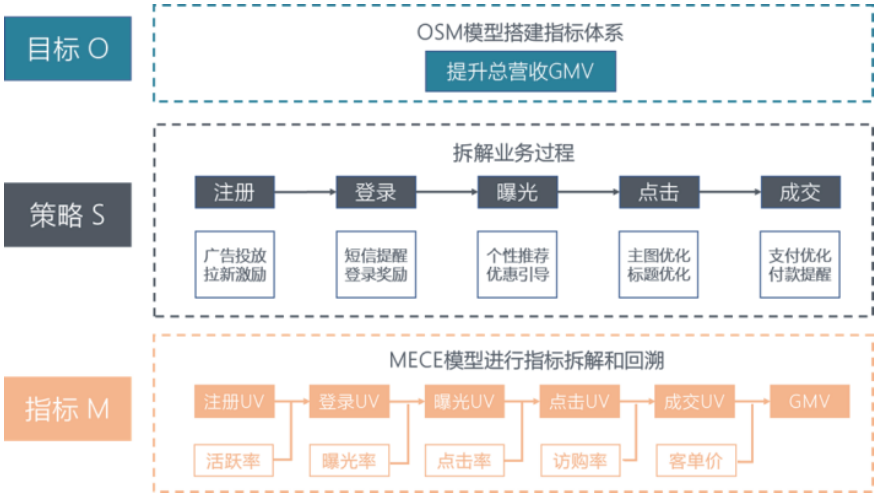
例如，针对第（2）步拆解的每个环节，建立对应的指标进行评估。在注册、登录、曝光、点击、成交各环节，可以通过各环节的 UV（Unique Visitor，独立访问数）去衡量。



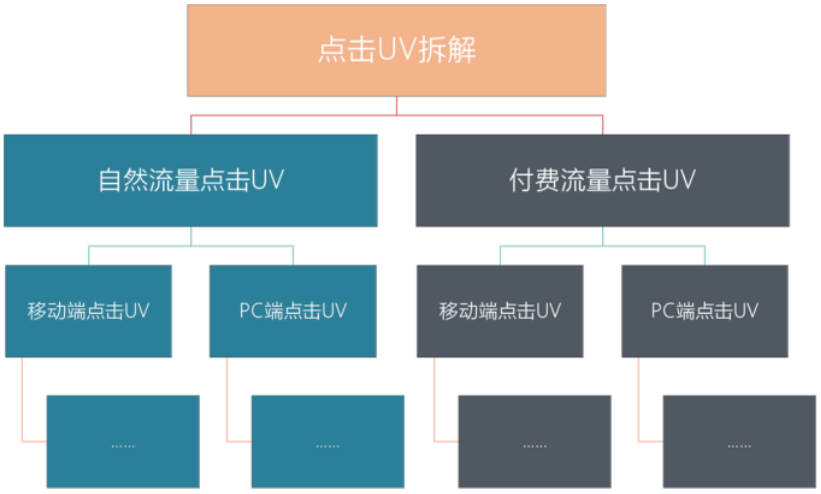

同时，还可以建立相邻环节之间的转化率，用于评估整个环节中各个漏斗的转化率，例如点击率（点击人数 / 曝光人数）用于衡量从曝光到点击环节的转化率。




经过以上一步步的拆解，最终形成初步的指标体系，如下图所示。



当然这个指标体系还比较简单，因为只进行了一层拆解，实际上针对以上每个过程，可以进一步拆解细分。例如，我们可以对点击 UV 按照来源渠道等进行逐层拆解，拆解成自然流量点击和付费流量点击，自然 / 付费流量点击又可以进一步细分为 PC 端和移动端的点击，以此类推，逐层拆解。




### 4. 小结

数据指标体系搭建的方式不拘一格，拆解方式也多种多样，但原则一定是结合着业务进行，因为数据指标体系最终一定是指导业务，帮助业务发现和解决问题，脱离了业务的指标体系只能是纸上谈兵，毫无意义。

## 第 8 问：数据指标体系如何应用？——数据监控体系

> 导读：数据指标体系指导业务，帮助业务发现和解决问题，那么实际工作场景中如何应用数据指标体系呢？此时，就需要配合“数据监控体系”，通过业务数据监控、分析、复盘等找出问题，寻求解决方案，为业务下一阶段目标进行预测和决策，有效地发挥出“数据指标体系”的作用。

### 1. 什么是数据监控体系?

“数据监控”即“采集 + 呈现”，也就是将用户全链路行为数据以及业务数据采集过来，并用可视化的图表呈现出来。“数据监控体系”就是将这些单一的数据指标体系与管理流程结合起来，来满足复杂的产品业务线的监控需求。

数据监控体系的重要性如下：

（1）反映过去产品和业务情况。

能够反映过去产品和业务的情况，对现在情况做对比和参考。

（2）现有业务线的状态监控。

对目前产品业务线的状态进行监控。

（3）发现数据异常等问题。

及时发现业务指标升高或降低，以及产生的原因。

（4）预测业务发展。

反映产品业务线未来可能发生变化的趋势，再根据指标数据控制成本等。

### 2. 数据指标体系的应用思路是什么？

（1）明确产品业务目标、KPI 和所处的产品阶段。

需要认清和明确目标（量化以及拆分目标是数据分析的灵魂）。一个业务目标的达成可能是多个团队、多个地区、多个渠道共同促成的，所以，在了解整体目标的同时也要关注局部目标，增加分类维度，明确局部的好坏状态。而判断业务走势正常还是异常，探索解决问题的办法，都是从计算目标和现状的差距开始的，这一点非常重要。不同的产品阶段有不同的产品目标业务。

目标细分可以有多种类型，常见的有以下几种：

- 按达成时间细分：年、季度、月。
- 按服务对象细分：各个部门、整个公司。
- 按流程位置细分：结果型目标、过程型目标。

（2）根据业务目标，确定判断标准。

依据判断标准，查看数据指标体系中的核心指标是否达标。没达标的话差多少，是亏空还是差一些，是什么原因造成的，问题大不大；达标了超出多少，为什么会超出，有没有更多的机会。判断标准的维度如下图所示。

**（3）根据业务需求，从数据指标体系中挑选相应数据指标，进行拆解。**

数据指标体系里有很多数据指标：日活（DAU）、月活（MAU）、下载量、激活量、新增注册量、次日留存率、次人均时长、首页访问率、停留率、人均充值金额（ARPU）、商品交易总额（GMV）、客单价，等等。

针对不同的指标，拆分不同的层级。不一定要拆得很细，否则层级会过深，基本上 3 个层级就能够指导我们去做一些动作。

**（4）查看不同层级的数据指标，找出原因。**

确定哪些数据指标没达标，是什么原因，是推广少、成本高、用户少，还是转化率低或者付费率低等。

**（5）搭建以日、周、月为单位的数据指标监控体系报表。**

监控每日、每周，以及上周、上月同周、上上月同周的数据报表，以图表展示，来反映产品的变化趋势，通过过去一周的数据反映产品现状，通过过去三个月的产品业务线数据变化趋势预估未来的变化趋势。

数据监控指标体系的基本逻辑是先看一级指标，再结合二、三级指标预测未来趋势。

**（6）根据数据监控结果 / 数据指标体系进行多维度分析，明确管理流程，实现控制。**

具体操作步骤如下：

①进行多维度分类分析。如：

- 哪些区域、团队、渠道，完成目标是下降还是持续上涨；
- 哪里没做好，是什么原因；
- 看看是谁能力大，是谁影响了整体。

② 确定指标异常状态，明确运营策略执行者。如：

- GMV 降了→客单价降低了→用户运营想策略；
- GMV 降了→某类商品降幅大了→商品运营想策略；
- GMV 降了→外部流量太少了 →渠道运营想策略。

③明确执行时间，要有时间状态和走向判断。如：

- 过去 + 负向→关注什么问题；
- 过去 + 正向 →发现什么经验；
- 未来 + 负向 →警惕什么风险；
- 未来 + 正向 →提示什么机会。

④明确需要多大力度。如：

- 注意出现异常；
- 提高、降低、保持等动作；
- 立即执行。例如，“客单价不能在 3 天内得到改善提高，本月 KPI 将不达标，要立即优化商品组合，提升客单价”。

⑤复盘改善后效果。最主要环节就是效果的复盘，而且要先看是哪个层数据指标的效果，再看具体效果并进行改善。

### 3. 小结

“数据指标体系”应用要配合“数据监控体系”，这需要我们不断地总结过往经验，了解未来产品业务计划，甚至收集一些竞品的情况，把整体现阶段的目标具体到某个人，有明确指向，不断地完善“数据监控体系”，发挥出“数据指标体系”的应用价值。

## 第 9 问：数据分析的产出价值是什么？

> 导读：随着大数据时代的到来，商业市场对“数据分析”相关岗位的就业需求也水涨船高。尽管如此，在与数据分析新人交流的过程中，仍会听到许多人对该岗位有“容易被替代”“发展前景窄”“价值难体现，沦为工具人”等疑虑。本问通过对数据分析价值的阐述，为心存顾虑、犹豫不决的读者提供一个参考。

在“第 2 问：数据分析是怎么来的？”这部分内容中，我们了解了数据分析的起源，明确了数据分析这个职能在经营全局中的定位。在此基础上，我们把镜头拉到微观的经营活动中，讨论数据分析是如何影响业务流程、产生价值的。

### 1. 描述现状——发现问题

有时业务可能并不存在确切的“问题”，需要通过加深对现有业务场景的理解、关键数据指标的监控（如每日新增用户数、DAU、转化率、复购率等），将数据可视化，用数据报告的方式呈现，来描述当下业务的现状，让业务相关人员对整体业务现状有所了解，以此来产出有效策略，优化业务现状。

例如，现在业务使用的是客单价平均值，将客户分为高、低两类人群进行营销，此时数据分析师通过对消费者进行洞察分析，给予更精准的人群划分方案：利用客单价分位数，将客户分为三类人群，这样业务利用更新后的策略进行营销设计，提高转化效果。分析过程可能是做相关分析、回归分析，甚至是无监督的聚类，来对现状进行解释，发现问题。

### 2. 解释原因——解决方案

通过数据发现某一指标异常的现象，需要进一步确定业务异动具体的原因。对产品或者用户行为中一些现象或者数据变化进行解释，让业务相关人员了解发生现象或者数据异常波动的原因，并针对性地给出解决方案。

例如，最常见的数据分析场景，就是业务相关人员发现销售额下降、用户流失、产

品跳失率高，也就是业务层面出现了一个待解决的问题，此时需要数据分析师介入，从数据层面挖掘原因、给出解决建议。分析过程可能是做一些探索性数据分析、统计分析、机器学习建模，甚至是做 AB 测试试验，最终交付分析报告，或者部署上线模型。

### 3. 总结原因——支持诊断

引起问题的原因是多方面的，要多方位思考，将关键指标逐层拆解，抽丝剥茧，从中找出问题的蛛丝马迹。此时，需要找到主要矛盾，让业务相关人员了解问题的根源，从而支持业务诊断。而支持诊断的内容主要集中在自动化报表，甚至是商业智能（Business Intelligence，BI）体系的搭建中。

例如，为了找出销售额低的原因，需要进行更多维度的拆分：销售额 = 客流量 × 客单价 × 转化率 × 复购率，要想分析销售额，就得从客流量、客单价、转化率、复购率这几个不同的维度去思考。数据分析师也可以通过交付“客单价预警报表”来优化该流程效率。

### 4. 进行预测——探索发现

业务中还存在一种需求，就是对未知的探索和预测。不同业务形态对需要探索和预测的指标不一样。社交类产品比较关注日活、新增等数据，电商类产品比较关注订单量、销售额、转化率等数据。而预测是对业务未来发展趋势的判断，有了精准判断可以让业务相关人员了解业务未来的走向，并制定针对性的防御措施（若进行深层次的建模，就要高层次的机器学习等技术作为支撑）。

例如，针对电商类产品，通过对比往年数据以及针对性的活动预期效果，在“双十一”“618”预测可能产生的流量的峰值，事先对服务器进行扩容，避免大流量冲击对业务造成影响。同时针对广告投放效果进行预测，有针对性地进行广告投放，确保流量，并根据数据指标的实时变化对投放进行动态调整。

### 5. 决策支撑——降本增效

所有数据分析产出的最终价值就在于指导业务决策，实现增长、降本增效。利用对比分析、描述性分析、多维分析、趋势分析等诸多分析方法对各种维度数据进行分析，挖掘数据潜在价值，为业务相关人员提供决策支持，提出解决方案，创造商业价值。

例如，可以利用数据分析筛选优质渠道。通过渠道分析，对比各个渠道新用户的留存，再结合各个渠道的推广费用算出 ROI，对比各个渠道费效比，筛选出优质、性价比高的渠道，从而加大在该渠道上的投放费用。

### 6. 小结

通过数据分析挖掘业务中的问题，并定位原因、给出方案建议，实现增长、降本

增效，是数据分析最重要的价值。但数据分析最终是否能产生价值，除了上述的发现问题、定位问题、给出方案和建议外，还要注重项目的落地，这里涉及的能力有**项目能力（需求管理、定义问题、落地计划、部署上线）、资源协调、向上汇报、横向沟通等**。

## 第 10 问：数据分析的常见陷阱有哪些？

导读：我们都知道数据驱动业务的时代，要拿数据说话。数据是反映业务、辅助决策的重要手段，但这些都建立在准确的数据分析结论上。在数据分析的过程中，尤其是对于刚入门的数据分析师，对**数据来源、统计口径、分析方法、业务经验、思考方式**等掌握不牢，很容易产生一些“陷阱”，以致分析的结论出现偏差。

本问将探讨在数据分析过程中几个常见的“陷阱”，给读者提供一些实用的经验，帮助读者在工作中规避这些“陷阱”。

### 1. 不了解数据来源，不确保数据的正确性

很多人在数据分析中十分重视分析方法，却忽略了数据本身，这是数据分析最大的“陷阱”：不了解数据来源，不确保数据的正确性。错误的数据是得不出正确的结论的，因此，数据分析的第一步就是了解数据来源，确保数据正确性。

例如，某 App 上线了新版的落地页，在不同渠道投放。数据稳定后，数据分析师发现某个渠道落地页的点击率、转化率等数据明显要好很多，建议加大这个渠道的投放。然而，突然接到研发人员的反馈，在数据埋点的时候不小心埋错了，这个渠道的数据是其他两个渠道数据的总和。因为错误的数据得出了错误的分析结论，还差一点做了错误的决策。

### 2. 未清洗数据，数据抽样出现偏差

梳理数据来源，确保数据的正确性是前提。但在正式开始分析之前，我们还要保证数据的质量和数据抽样的合理性，少数脏数据和异常值可能会使分析人员得出相反的结论，不合理、不均匀的抽样也可能使得分析结论与整体情况背道而驰。

由于程序错误、第三方攻击、人为等原因，数据采集中很容易出现极端异常值、缺失值等情况，这些脏数据会对分析结论造成很大影响，所以在进行数据分析之前，需要检查各字段的空值、数据分布等情况，并进行异常值剔除、缺失值填充等处理，保证数据质量后再进行分析。

另外，如果受限于数据样本量，要从总体样本中抽样进行分析，想要保证群体样本能够代表整体，就要保证样本均匀随机，避免人为主观的选择性偏差导致结论的偏差，进而得出真实可靠的结论。

例如某 App 升级后，想通过新版本和老版本用户的活跃情况对比，判断新版本是否优于老版本，但这里实际就隐藏了选择性偏见，升级新版本的用户往往本身就是较为活跃的用户，其活跃情况大概率优于未升级用户，这就是分析样本导致的结论偏差。

### 3. 需求不匹配，分析目的不明确

在了解数据来源并确保了数据质量后，接下来就要明确业务方真实的需求，问题到底是什么，明确了这个才能明确分析的目的，然后针对分析目的，搭建分析框架，选择分析方法，抽取特定的数据进行分析。避免为分析而分析的误区，才能得出正确而有价值的结论。

例如，某 App 的产品经理觉得目前产品转化率较低，想让数据分析师进行分析。如果数据分析师没有进一步确认是哪些用户 / 哪个环节转化率低，就开始拉取数据进行分析，很容易乱撞一通抓不到重点。其实产品经理说的是“新用户成单”的转化率低，明确了这个分析目的，数据分析师才可以继续分析是新用户来源不精准，还是引导不够等。

### 4. 指标不合理，评估出现偏差

明确了分析目的，下一步就需要选择合适的指标去定量评估问题，也就是要定义合适的数据指标。每个指标都有特定的统计逻辑，反映事物某一方面的特点。因此，在进行数据分析时，如果指标定义不当，很容易得出错误的结论。

例如，我们经常使用平均值来描述一组数据的集中趋势。但是，有些场景并不适合使用平均值，如果把我和世界首富的财富取平均值，我也是富翁，但很明显，这个平均值没有任何意义，因为个人财富并不服从正态分布，使用分位数、加权平均数可能更有意义。

### 5. 轻视业务，生搬硬套方法论，与实际场景脱节

定义好合适、准确的数据指标后，接下来就是使用各种数据分析方法来分析数据，得出结论，辅助业务决策。数据分析方法论是对一个数据分析项目起到指导作用的思路框架，掌握一些常用的分析方法论可以帮助我们高效地开展分析工作，但实际工作中切忌生搬硬套，不同行业、不同业务、不同阶段，适用的分析方法都有所区别。

例如，同样是用户分析，To B 业务和 To C 业务就有很大的区别，To B 的产品一般是解决系统化、流程化问题，关注更多的是效率、功能性等。而 To C 产品则偏向用户体验，要做到让用户“爽”，更多关注的是用户痛点、产品的交互和使用习惯等。所以，在数据分析过程中，不能完全生搬硬套历史案例的分析方法，而应重视对业务的理解。

实际业务往往比数据更加复杂，分析时需要了解具象化的业务场景，而不只是抽象的数据。数据分析师极易犯的错误就是只懂工具，没有真正理解业务需求。数据分析师

一定要多去一线了解业务，多站在业务的角度思考问题，想他人所想，急他人所急，这样才能及时甚至提前帮助业务方解决各种问题。

同时，数据分析师还要及时与业务方沟通，共享数据分析的成果，及时吸取业务方的反馈，不断地更新迭代分析结果，完成“从业务中来，到业务中去”的完整闭环，这样才能体现数据分析的真正价值。

### 6. 小结

以上都是工作中常见的一些数据分析“陷阱”，随着大数据时代的到来，数据量急剧增长，业务场景和数据分析也变得越来越复杂，我们要抱着敬畏的态度，谨慎地使用数据，大胆假设，小心求证，确保数据分析每个环节的可靠，避开数据分析中的“陷阱”，做出准确而有价值的数据分析。

## 第 11 问：如何让数据驱动业务？——数据分析流程

> 导读：将决策方向从“业务经验驱动”向“数据量化驱动”转型，可以更好地管理企业、驱动业务线的改进、挖掘业务的增长点。本问从业务解决问题的流程出发，探讨关于数据驱动业务增长的底层逻辑。

虽然我们一直说数据赋能业务，但是这个过程是如何落地的呢？回到业务层面，借助黄金思维圈模型，可以把解决问题的过程抽象成**明确问题（What）→ 分析原因（Why）→落地执行（How）**，贯穿这个过程，数据驱动的逻辑可以对应为**业务解构 → 建模分析 → 变革提效**。


### 1. 明确问题

何为问题？问题是当前课题下，现状与预期之间的差距。

业务场景下，所见即为现状，例如本月销售额 100 万元，消费人数 1 万人，客单价 100 元，这些都是数据现状。一旦预期与现状不符，例如本月预期目标销售额 300 万元，消费人数 1.5 万人，客单价 200 元，这中间的差距（销售额 -200 万元，消费人数 -5 千

人，客单价 +100 元）就是问题，这里的“问题”是广义的：针对表现不好的地方，提出解决方案；针对表现好的地方，提炼成功经验。这些都是分析的起点。

业务问题的产生往往依据的是直观的结果：本月销售额不达标、领导要求提升客单价等，而这些原始的需求往往是模糊且无从下手的。只有明确真正的问题是什么，才能解决它。

因此，在该阶段，**数据真正发挥价值的地方在于通过“业务解构”定位核心问题点**。我们说数据分析一定是从业务出发，最终再回到业务落地的过程，而明确问题阶段，对应的就是“从业务出发”：在业务场景下，定义业务现存的问题，或者明确业务期望此次分析能达到的目标，只有明确了目标，才能给分析过程带来明确的方向。

但对业务问题的定义又不能局限在业务层面，之所以是数据分析，是因为需要借助数据的力量，把业务问题转换成数据分析需求，这样才能应用“武器库”里的分析方法解决问题。也就是说，在明确问题阶段，我们需要做以下事情：① 从业务层面明确问题；②将问题转换成具体的数据分析需求。

这里有一个问题拆解的逻辑，业务方给数据分析师的问题大多是笼统定性的，容易让人无从下手。该怎么办呢？解决方案就是上述的② ：从数据层面定义（或拆解）业务问题，把大问题拆解成可解决的小问题，然后在分析原因阶段一个一个解决。

例如，逻辑树方法（详见第 25 问）就将需求来源的业务拆解，根据生意公式：销售额 = 消费人数 × 客单价，把销售额指标的变动拆解到更具体的层面。假设销售额（−30%）= 消费人数（−40%）× 客单价（−30%），此时可以能看到，面对销售额不达标的场景需求，真正需要解决的是消费人数下降的问题。

### 2. 分析原因

在明确问题阶段，我们能得出此次数据分析项目要达成的分析目标，以及从大问题拆解而来的小问题。除了解决这些问题外，在分析原因阶段有一个重要的任务：追溯数据变动的原因。只有“知其所以然”，才能扩大成功经验、汲取失败教训。

在以往，很多情况下仅能依据业务经验解决问题。这种反馈形式，有如《思考，快与慢》中“系统一”的快思考，建立在个人业务经验基础上，很容易产生偏见。当然不排除存在经历大量刻意练习，或者有丰富实战经验的专家，仅凭感性的认知就能做出正确决策。但是对于大部分人来说，直觉支撑的决策往往站不住脚。

因此，更需要《思考，快与慢》中“系统二”的慢思考，通过深思熟虑的分析、验证，寻求解决问题的方案。数据分析提供基于数据支撑的框架思考能力，就属于这种反馈形式。

**在该阶段，数据发挥价值的地方在于“建模分析”**。例如为了解决上述“消费人数下降的问题”，可以搭建 AARRR 漏斗模型（详见第 33 问），通过提升上游留存（Retention）阶段的人数，进而提升消费人数（Revenue）。假设对用户行为数据分析发

现，只要用户邀请超过 10 名好友，用户 30 天留存的概率就会从 30% 提升到 70%，因此，就能给出解决问题的业务策略：通过分享游戏刺激用户邀请超过 10 名好友。在这个案例中，数据分析通过找到业绩提升的“魔法数字”驱动业务增长。

当然，这仅是其中的一种分析框架、一种驱动路径。概括来说，帮助驱动业务的数据分析方法可以概括为四种：**比较分析、相关分析、预测和发现**。

**（1）比较分析。**

指标的好坏、特征是否显著等都可以通过比较分析的方法来实现，例如常见的归因业务场景，本质就是做比较，通过横向、纵向的比较找出原因。

分析方法：T 检验、方差分析、同比、环比、同期群分析等。

**（2）相关分析。**

分析变量之间的相关性是重要的分析场景。例如，业务中想知道提高广告预算能否或者能提升多少销售业绩，运用相关性分析或许能找到最优投放 ROI 的配置方案。

分析方法：卡方、皮尔逊（Pearson）相关系数、斯皮尔曼（Spearman）相关系数、结构分析等。

**（3）预测（有监督）。**

不论是对企业销售的预测，还是对用户行为的预测，都能帮助提升业务效率。例如常见的预测用户流失分析，得到高概率流失的人群名单后，运营及时通过提前营销干预，提高用户留存率。常见的销售预测则能帮助企业在供应链侧做准备。这类场景主要应用的是机器学习中的有监督分类模型。

分析方法：线性 / 逻辑回归、决策树、时间序列分析、贝叶斯等。

**（4）发现（无监督）。**

前面三种分析方法都是基于企业已知模式的分析逻辑，还有一种分析方法——无监督的机器学习模型，可以应对未知模式的分析。例如，不知道应该把现有人群分成多少个组来进行营销最合适，就可以对人群基于核心特征做无监督的聚类分析，得出有效分组的界限。

分析方法：Kmeans 聚类、DBScan 聚类等。

接下来的第 2 章会为读者带来这些方法更多、更具体的介绍。

### 3. 落地执行

至此，我们在分析原因过程中得到一系列的数据结论，这些数据结论最后还需要通过结合业务场景的定性分析形成业务结论。在业务结论的基础上，我们才能给出落地的业务建议。

什么叫“落地”？即保证业务方可以操作且愿意执行。

“可以操作”说明所给的建议的颗粒度足够细。例如，“要提高客单价”就不可操作，业务方不知道要怎么做才能提高客单价，而“通过促销活动提高某产品系列销售占

比至 40%”就可以操作，业务方马上就可以围绕该产品系列给出方案。

“愿意执行”说明所给的建议是符合业务方利益的。实际工作中，每个部门都有不同的工作，假设我们给用户运营部门提了产品优化的建议，那用户运营的同事也无法马上对产品做任何操作，因为这超出了他们的权限。所以，要给用户运营部门提用户活动相关的建议才是正解。

面对业务问题，不论分析的过程是复杂还是简单，最重要的是要做出行动，完成数据驱动的“最后一公里”，**在该阶段，数据发挥价值的地方在于“变革提效”**。借用《数据分析即未来》里的观点：判断一个组织的数据能力强不强，并不在于它的算法模型有多复杂，而是数据模型能否融入业务流程中，在不同部门间形成协同。为了达成数据驱动过程，在最后的落地阶段，需要数据分析师完成两项工作：数据故事与模型实施。

**（1）数据故事。**

分析项目的落地需要多方参与，即使是业务实践经验丰富的分析师，由于流程边界的存在也不可能每步都参与执行。因此，确保项目有效落地的一个必要条件是和业务方达成共识。

为了与业务方达成共识，需要讲数据故事，阐述起因（需求定义）、过程（分析逻辑），确保结局（重要结论）引人入胜（被认可）。这个过程需要制作 PPT 向上汇报、与业务方沟通，甚至是做跨部门的演讲。

**（2）模型实施。**

不论是业务模型还是算法模型，最终都需要落地实施、部署上线。到这一步，数据分析结论对业务流程则会产生实质性的影响。

- 对于业务模型，如 RFM，则是部署到业务流程中，应用在会员管理、活动营销等环节；
- 对于算法模型，如推荐算法，则是部署到产品功能上线，可以通过内置算法、REST 接口等形式落地。

### 4. 小结

数据驱动业务增长是一个厚积薄发的过程，需要在日常业务工作中做好数据收集、数据清洗、数据监控、数据可视化分析、数据产出在内的每一个环节。其底层逻辑体现在基于数据思维进行的业务解构、建模分析，并最终将分析结论在业务流程中落地，实现变革提效。