第一篇

数据分析的方法

数据分析已然成为当下最热门的运营技能，大有“不会数据分析都不好意思和别人打招呼”之势。回顾我十多年大数据工作中所经历的行业，包括通信企业、智慧城市运营商、国内顶尖互联网公司以及金融企业，他们都对数据分析有着较高和迫切的要求。特别是近些年参加行业论坛、互联网分享会和开发数据分析培训课程时，能感觉到大家对于数据分析有着很旺盛和迫切的需求，也能明显感觉到大家对于数据分析有一些共性的疑问：

- 如何入门数据分析？
- 如何掌握有效的数据分析方法？
- 如何成为厉害的数据分析师？
- 数据分析一定要会 Excel、SQL 和 Python？

我相信这些问题也代表了大多数读者的疑问。事实上，学习数据分析有一套非常科学的方法。这个学习方法要求我们首先掌握一种数据分析的流程、思路和方法，并学习这个流程、思路和方法由哪些步骤组成，每个步骤用到了哪些分析技术，以及这些分析技术的适用场景是什么。在掌握这个数据分析流程、思路和方法的基础上，我们再寻求一个合适的数据分析工具来实现和执行这些流程、思路和方法。

这种学习方法的好处有以下几点。

第一，掌握一个科学的分析方法之后，再寻求一种合适的分析工具，可以让我们分析数据的效率大幅度提升。

第二，避免出现本末倒置，也就是学完语言，学完算法，具备一定的编码能力后，回到工作岗位中依然不知道从何下手，在学习技能和应用技能之间出现了巨大的断层。出现断层的原因就是我们虽然学习了工具，但是没有掌握应用工具的思路和方法。

基于此，本篇为大家带来一套有趣、有效、有料的数据分析方法。

首先，这套数据分析方法适用性非常广，不仅适用于互联网产品，也适用于线下业务，而且这套数据分析方法难度非常低，所应用的知识也只是大学期间统计学中的部分基础内容，可谓“简约但不简单”。

其次，选择 Excel 作为这套数据分析方法的落地工具，原因在于我国 Excel 普及率非常高，可以说 99% 的公司和个人都在使用 Excel。并且，在

Excel 中实现这套数据分析方法，绝大多数情况下只需点点鼠标即可完成，极个别特殊场景才需要更高级的函数来辅助，所以特别适合产品、运营、市场、营销、销售、管理等从事业务运营的读者，基本上可以做到 3 分钟即了解，5 分钟即掌握，10 分钟即熟练。

最后，每个数据分析的方法都列举了实际工作和工程应用案例，通过实际案例的拆解分析，让读者更有代入感和共鸣，尽可能降低学习与应用之间的门槛，真正做到即学即用。

数据分析的完整流程包括 3 个主要步骤，分别是：

- 寻找并准备数据：如何收集、处理与清洗数据；
- 从数据中寻找问题的答案：如何进行数据分析与建模；
- 用分析支撑决策：如何从数据中洞察业务并输出结论。

先说说这个流程的特点：

闭环，从业务中来，到业务中去。数据分析的结果可以在闭环中落地执行，在执行中验证效果，并执行新一轮的闭环分析。

通用，普适。从上述步骤的描述上看不出与任何行业、产品相关的词汇，意味着这几个步骤是行业通用和业务普适性的。既可以用这套分析方法分析电商产品的商品运营、供应链运营、渠道运营、品牌运营和用户运营，也可以用这套分析方法分析内容产品的内容消费情况，以支撑内容的热点运营、平台运营、品牌运营等。

在详细拆解每个步骤的内容前，我们先快速概览一下这 3 个步骤的主要特点以及使用场景。

1. 寻找并准备数据：如何收集、处理与清洗数据

寻找并准备数据，主要阐述数据预处理工作。正所谓磨刀不误砍柴工，通过数据预处理，我们抹除脏数据、移除空白数据，将数据格式统一，目的是在提升数据质量的同时规范所有的数据指标，以方便后续分析，降低分析难度，提升分析速度。

例如，原始数据中日期格式是日－月－年，而我们的要求是年－月－日，两者格式并不统一，必须通过数据预处理进行格式转换。

例如，原始数据中存在空白值和特别大或特别小的值，不做预处理的话会导致分析结论出现偏差甚至错误，所以必须进行数据预处理。

在这个部分，我们用 Excel 进行数据预处理。

2. 从数据中寻找问题的答案：如何进行数据分析

在数据预处理之后进入最主要的分析步骤，即从数据中寻找问题的答案。在这里将阐述 5 种数据分析的方法，即用描述性统计寻找数据整体和表象特点，用变化分析寻找数据分析的切入口，用指标体系来寻找变化的原因，用相关性分析判断原因的影响程度，用趋势预测来分析数据未来的发展趋势。

这 5 把利器既可以按照顺序使用，也可以拆解出来单独使用，由此体现了这套数据分析方法的灵活性所在。

在这个部分，我们依旧用 Excel 来实现这 5 把分析利器的作用，而且只需要掌握 Excel 的基本操作即可，不需要 VBA、函数等高级技能。

3. 用分析支撑决策：如何从数据中洞察业务并输出结论

通过上述数据分析武器寻找出来指标数据背后的原因以及发展趋势之后，还需要进一步将结果从数据转化为运营策略。在这里提出一种 Business-Operation 模型，借助 Business-Operation 模型将数据分析结果转化为可落地的运营策略。

注意：数据报表和数据分析报告的技巧不在本书中展开，请同学们自行搜索学习。

第 1 章 准备工作：数据清洗与预处理

本章介绍了数据预处理与清洗的原理和流程，并通过 Excel 完成常见的数据预处理和清洗操作。数据经过预处理和清洗后才能被高效分析和挖掘。

本章涉及的知识点：

- 数据预处理的流程
- 用 Excel 实现常见的数据预处理

1.1 为什么要正确和高效地预处理与清洗数据

本节首先介绍了数据的加工和生产流程，并在此流程中详细分析数据预处理的 3 个步骤，以及完成每个步骤的具体方法。

1.1.1 指标的数据来源

开始拆解数据预处理前，非常有必要和大家聊一聊数据的加工和生产流程，因为：

（1）数据的加工和生产流程是数据分析的基础。

（2）可以了解指标数据是如何从业务系统一步步汇总计算，从无意义的明细数据变成具备业务意义和价值的指标的。

（3）可以快速分析数据问题，便于后期快速定位和追查数据问题。

如图 1-1 所示，要生成我们日常运营的指标，数据需要经过至少三大节点，即源系统、数据中台和数据应用层。

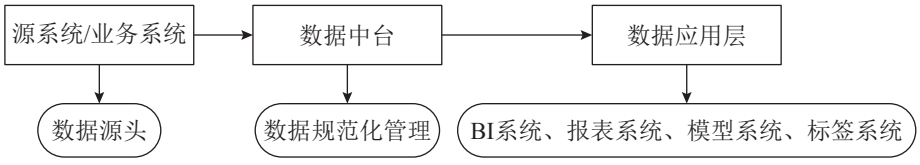

图 1-1 数据的生产流程

1. 源系统

源系统通常也叫业务系统，即承载产品业务的系统，它们在运行各种业务应用的同时也会产生对应的业务数据，所以叫源系统，是数据产生的源头。

特别对于平台类产品，因其承载了很多业务，故而连接了很多业务系统。令人苦恼的是，这些业务系统都是由不同供应商开发和运维的，各个源系统间的数据结构不完全兼容，不仅数据字段不一样，甚至同一业务含义字段的命名、格式、约束也不一样。

例如，源系统 A 中用户标识用的是手机号，源系统 B 中用户标识用的是注册用户名，源系统 C 中用户标识用的是微信 ID。

如图 1-2 所示，对于用户本身而言，手机号、注册用户名、微信 ID 等都是这个用户的唯一标识，但是在不同的源系统中却是完全不同的标识。显然，如果把手机号、注册用户名、微信 ID 作为 3 个用户是违背认知且不合理的，需要某种机制把这 3 种类型不同但却是指向同一用户的标识统一起来，让数据认为他们是 1 个用户，而不是 3 个用户。

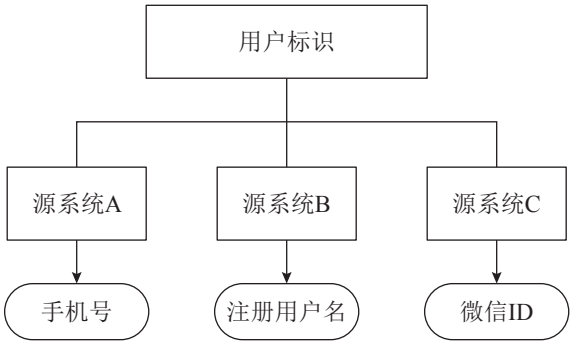

图 1-2　不同源系统用户标识不尽相同

例如，源系统 A 和源系统 B 中用户标识用的都是手机号，但 A 中的手机号格式是 139-1234-1234 的 3-4-4 结构，B 中的手机号格式是 139-123-41234 的 3-3-5 结构。

如图 1-3 所示，对于用户本身而言，这两种格式的手机号都是 13912341234，都表示了同一用户，区别仅仅是存储的格式不同，但是如果不做数据预处理，就会被认为是两个用户，显然不合理。

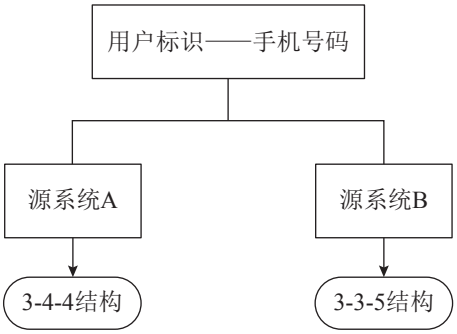

图 1-3　不同源系统同一字段格式不尽相同

我们把上述情况叫作异构系统中的数据规约问题。

异构系统中的数据规约问题，不仅会带来数据的混乱，还会造成大量的数据孤岛，导致有用的数据无法被发现，故而需要在数据预处理阶段将同一业务含义的不同数据整齐划一，以同一种数据规范进行存储和分析。

注意：异构系统的数据规约是数据治理的重要工作。一般平台型 APP，后端都连接超过 300 个业务系统，这些业务系统区别极大，让异构系统的数据规约成为极其重要的工作。这种情况下，通常会有一个中等规模的团队来负责 300 多个业务系统的数据规约工作。

2. 数据中台 / 数据中间层 / 业务中间层

数据中间层，也叫业务中间层、数据中台，其作用是将源系统中异构系统的数据统一规范和统一管理。一般有两种方式：实体整合和逻辑整合。

（1）实体整合

实体整合，即 Entity Integration，定期将源系统的数据抽取到数据中间层，并在数据中间层完成数据预处理，以保证数据进入数据中间层后已满足统一规范和统一管理的要求。

实体整合的第一个特点是非实时，即数据中间层的数据与源系统的数据存在时间差，这个时间差通常用 $T+x$ 来表示。

实体整合的第二个特点是数据已经过预处理，对于上层数据应用和业务分析而言速度很快，且不受源系统干扰，源系统即使中断服务或停服维护也

不影响分析，因为数据已经被抽取到数据中间层。实体整合特别适合源系统更新不频繁的数据，例如用户信息、商品信息等，以及时效性要求不高的指标，例如周报、月报中的指标。

注意：*T*+*x*，即数据中间层数据的更新速度比源系统落后 *x*，通常单位为天。*T*+0，即数据中间层的数据与源系统保持同步；*T*+1，即数据中间层的数据比源系统要落后 1 天，意味着在数据中间层看到今天的数据实际上是源系统中昨天的数据。

（2）逻辑整合

逻辑整合，即 Logical Integration，不需要定期将源系统的数据抽取到数据中间层，不在数据中间层完成数据预处理，而是创建某种映射关系，连接数据中间层和源系统，并在分析数据的时候实时进行数据预处理。

逻辑整合的第一个特点是实时，即数据中间层的数据与源系统的数据是映射关系，不存在时间差，是 *T*+0。

逻辑整合的第二个特点是数据未经预处理，对于上层数据应用和业务分析而言需要在数据使用过程中实时进行预处理和计算，速度会受影响，且受源系统服务质量的影响。如果源系统中断服务或停服维护，将无法完成数据预处理和分析，因为数据是实时连接到源系统的。

逻辑整合特别适合源系统更新频繁的数据，例如交易明细、浏览记录等，以及时效性要求较高的指标，例如用户访问路径分析、用户即时兴趣推荐等。

说明：逻辑整合，类似我们在 Windows 中为文件创建快捷方式，在 Linux 中为文件创建 soft-link。

无论实体整合还是逻辑整合，其整合方式最常见的是统一用户标识，即通过某种用户标识来将各个业务系统的数据整合，形成复合的星型结构。统一的用户标识既可以是具备实际意义的信息，例如手机号、用户名等，也可以是数据中间层自行创建的标识，只要保证唯一性即可，最常见的是将用户个人信息进行不可逆加密来生成唯一标识的字符串，既保证用户隐私，也能唯一标识用户。

例如，阿里的数据中台通过 OneID 来连接用户在阿里巴巴产品体系下的用户数据，将用户在阿里系产品中产生的数据通过 OneID 关联起来，这样就可以从衣食住行等各个方面来精确描绘用户画像和兴趣，制定高转化率的营销方案。

注意：个性化推荐系统分为离线更新和在线更新。离线更新通常在用户行为热度降低的时候进行，一般选择在晚上更新。在这个时间段可以分析用户过往很长周期——例如一年——的数据来计算用户的兴趣，从而完成内容推荐。在线更新通常在用户使用产品过程中进行，计算用户的即时兴趣，实现实时推荐。

3. 数据应用层

数据应用层通常包括我们熟知的 BI 系统、报表系统、模型系统、标签系统等，它们都是基于数据中间层的数据应用服务。数据应用层也是数据分析人员最为频繁接触和使用的地方，在这里数据分析人员利用管理驾驶舱、各种业务报表等来分析业务变化，寻找原因，制定策略。

注意：在入职新公司的数据分析岗位时，第一件事就是向同事详细了解数据生产和加工流程，即指标数据来自哪个业务系统，经过哪些中间系统进行汇总计算，最终在报表层面如何体现。了解指标的加工流程，对于后续指标运营和分析有极大的帮助。

1.1.2 数据预处理的目的

上面讲了指标数据如何从源系统最终达到数据应用层，其间多次提到数据预处理，那么数据预处理的目的是什么呢？

数据预处理的目的是在对业务数据进行分析挖掘前，先行对数据进行一些处理，以提升数据质量，为数据分析过程节约时间和空间。

注意：在实际工作中，数据预处理的时间甚至要超过数据分析本身所需的时间。如果需要分析的数据质量很高，那么对后续的数据分析来说无疑是如虎添翼。

1.1.3 数据预处理的流程

在数据科学领域，已有一套成熟的数据预处理流程，并在实际工程中应用多年，成为实际业务运营中的标准，它就是 ETL，即 Extract（抽取），Transform（转换）和 Load（加载）。

ETL 是一种数据预处理流程，它负责从异构的源系统中读取数据（所谓的 Extract，即抽取），然后根据一定的数据处理规则进行数据清洗和转换（所谓的 Transform，即转换），最后将预处理完成的数据加载到数据中间层或数据应用层（所谓的 Load，即加载）。

如图 1-4 所示，方框部分就是 ETL 部分。ETL 向下对接源系统，向上对接数据中间层或数据应用层。所有的数据预处理工作都在 ETL 中进行。

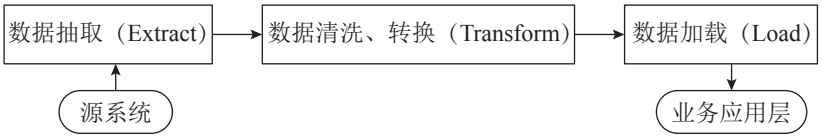

图 1-4　数据预处理流程 ETL 的架构

下面简要说明 ETL 的 3 个模块，更多信息请参详 ETL 的专业书籍。

1. 抽取（Extract）

抽取，即从源系统中读取原始业务数据，常见的数据源以关系型数据库（Oracle、PostgreSQL、MySQL、SQLServer）、分布式文件系统（Hadoop、Hive）为主。

抽取数据主要有两种方式：增量抽取（Incremental Extraction）和全量抽取（Full Extraction）。增量抽取，即每次从源系统抽取数据时仅抽取更新的数据，包括新增、更新和删除的数据，无变化的数据不会更新；全量抽取，即每次从源系统抽取全部数据，包括新增、更新和删除的数据，也包括无变化的数据。

2. 转换（Transform）

转换，即根据数据中间层或数据应用层的要求，将从源系统抽取的数据进行转换处理，主要包括数据清洗和数据转换。数据清洗，是指清洗掉重复的、

不完整的以及错误的数据；数据转换是指按照预处理规则将源系统中的数据转换为符合规范的数据格式。

常见的数据转换策略包括如下内容。

（1）移除非业务列

移除非业务列，即将与业务分析无关的列全部删除。常见的非业务列包括自增序号列、标识 ID 列、预留的空白字段列等，如图 1-5 所示。

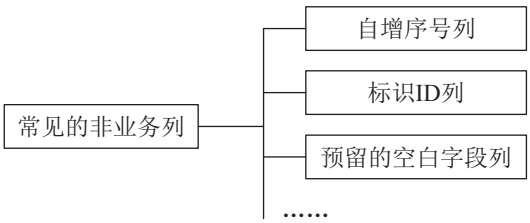

图 1-5　常见的非业务列

（2）重复值处理

重复值，通常是由于数据抽取过程中未进行排重判断，或源系统中数据创建时发生错误。重复值通常需要删除，多条重复数据仅需保留一条。

（3）缺失值处理

缺失值，通常是由于源系统服务故障无法上报数据，或源系统进行迁移暂停服务导致无数据。根据业务运营的需求，可以选择保留缺失值、回补缺失值和删除缺失值 3 种策略。

在实际工程应用中，最常见的缺失值处理策略是借助缺失前后的数据进行均值回补。

某 APP 用户行为分析系统，在 2019 年 5 月到 6 月进行系统迁移，导致其间用户行为数据缺失且无法后补，可以借助 5 月前和 6 月后的正常数据趋势来回补 5 月至 6 月的数据，使整年度的行为数据趋势完整。当然，这需要在系统中清晰标注数据回补的逻辑。

（4）文本数值化

文本数值化，即将文本存储的数据转化为数值，以方便后续的汇总计算。文本数据仅能进行计数汇总，无法满足后续复杂的数据分析。

常见的文本数值化策略包括：性别数值化，男 - 女转为 1-0 或 0-1；地理信息数值化，直辖市为 0，省会为 1，或一线城市为 1，二线城市为 2；以文本存储的数字，转换为以数值存储的数字。这个是在数据预处理中易被忽略的情况，极易导致后续汇总分析发生错误，特别需要仔细处理。

（5）数据离散化

数据离散化，即将连续的数据按照规则转换为离散的数值，有时也叫分桶转换，即按照设定的分桶规则，将数据转换为分桶的信息。通过数据离散化，可以将连续的数据转换为更接近于业务运营角度的表述，在后续运营中更加直观。

数据离散化的方法主要有 3 种。

a）等宽法

等宽法，即将连续的数据分为长度相同的多个区间，每个区间赋予一个业务定义。

例如，用等宽法将用户年龄数据进行离散化。

用户年龄数据：1，9，14，18，20，32，37，40，48。

第一步：设定区间宽度为 10，每 10 岁分到相同的区间。

第二步：根据年龄的最小值和最大值，结合区间宽度即得到 5 个区间：[1,10]，[11,20]，[21,30]，[31,40]，[41,50]。

第三步：为这 5 个区间设定业务名称，即 A：[1,10]，B：[11,20]，C：[21,30]，D：[31,40]，E：[41,50]。

第四步：将用户年龄按照区间范围和名称进行离散化，如表 1-1 所示。

表 1-1　等宽法进行数据离散化

1	9	14	18	20	32	37	40	48
A		B			D			E

于是，用户年龄由 9 个值转换为 4 个值，完成离散化处理。

b）等频法

等频法，即将连续的数据分为个数相同（频次）的多个区间，每个区间赋予一个业务定义。

例如，用等频法将用户年龄数据进行离散化。

用户年龄数据：1，9，14，18，20，32，37，40，48。

第一步：设定区间所含数据个数为4，即每4个数据分到一个区间，若区间中数据个数为4，则下一个数据自动分入下一个区间。

第二步：将用户年龄按照区间频率进行离散化，如表1-2所示。

表1-2　等频法进行数据离散化

1	9	14	18	20	32	37	40	48
A				B				C

于是，用户年龄由9个值转换为3个值，完成离散化处理。

不过在实际工程应用中，等宽法和等频法并不常用，原因在于以下三点。

第一，等频法和等宽法都是根据数据分布特点来进行离散化，未将具体业务需求考虑在内，数据离散化处理后可能难以在业务运营中落地。

第二，等宽法可能会将相同的数据分到两个区间。例如原始数据中第10位和第11位数据都是10，在区间宽度为10的条件下，第10位和第11位的两个数据就分到两个区间了。如果这个数据表示的是年龄，显然不应该分到两个区间。

第三，等频法可能会将差异较大的数据分到一个区间。例如原始数据前三个数是10，12，14，第四个数是50，在区间频次为4的情况下，10、12、14和50就分到一个区间了。如果这个数据表示的是年龄，显然不应该包含50。

所以，在实际工程应用中，更常见的是用自定义区间法进行数据离散化处理。

c）自定义业务区间法

自定义区间法，即按照业务运营的要求来设定离散化的规则，包括离散化区间的个数，区间的宽度，以及区间对应的业务含义。假如源系统中有一个年龄列，记录了用户的年龄，但在业务运营中不会直接应用具体的年龄数值，而是应用少年、青年、中年和老年等年龄段标识。这时就需要按照年龄段标识的规则将具体的年龄数值转换为少年、青年、中年和老年，如表1-3、表1-4和表1-5所示。

表 1-3　源系统的数值化年龄

用户	A	B	C	D	E	F	G
年龄	11	32	19	64	49	37	26

表 1-4　业务运营需要的年龄段标识规则

标识	少年	青年	中年	老年
规则	(10 , 20)	(21 , 30)	(31 , 50)	(51 , 100)

表 1-5　转换后的年龄段标识

用户	A	B	C	D	E	F	G
年龄	少年	中年	少年	老年	中年	中年	青年

（6）数据归一化 / 标准化

数据归一化 / 标准化是将不同的数据转换到同一范围或同一标准，以方便后续的分析，主要包括两部分工作。

a）将相同量纲但不同范围的数据缩放到相同的范围，以方便对比分析

例如分析班级中语文和数学的考试成绩，发现语文成绩的平均分是 110 分，数学成绩的平均分是 90 分。此时可以断言语文成绩就比数学好吗？

两者量纲虽然都是得分，但语文的成绩范围是 0 ～ 150 分，数学的成绩范围是 0 ～ 100 分，两者的成绩范围不同，即评判的基准不一样，显然不能断言语文成绩就比数学好。

若要科学对比这两门课程的成绩，就需要将两门课的成绩范围缩放到同一区间，使两门课程的评判标准一致，这样才能进行成绩好坏的判断。

又例如分析下沉市场，或所谓的“五环以外”的非一、二线城市的用户购买力和可支配收入时，显然不能拿四、五线城市的用户和一、二线城市直接进行对比。原因在于，四、五线城市和一、二线城市的居民收入、消费水平和生活成本等数据指标的范围完全不同，在四、五线城市一顿午餐可能只要 10 块钱，而在一线城市一顿午餐可能需要 20 块钱。

如果直接进行对比分析，极有可能得出错误的结论：四、五线城市的用户购买力和可支配收入远远低于一、二线城市，亦有可能制定出错误的策略：暂时不进入四、五线城市的下沉市场。

如果将四、五线城市和一、二线城市用户的收入、消费和生活成本等数

据指标的范围缩放到同一标准，就能进行公平的对比分析，或许得出的结论是四、五线城市的居民购买力和可支配收入并不低于一、二线城市，可以快速进入四、五线城市的下沉市场。

事实上，很多调查机构也都做出过这样的结论：小镇青年们的购买力惊人，幸福指数远高于一、二线城市用户，就像拼多多、快手、抖音等都是把握住了下沉市场的风口，迅速成为中国移动互联网的重要流量平台。

b）将数据的微小变化充分放大，以方便描述分析

在内容运营中 CTR（点击率）是重要的指标，代表了内容（短视频、图文）的质量以及受欢迎程度。CTR 是一个用百分比率表示的指标，范围从 0% 到 100%。在实际运营中，由于马太效应而导致大部分内容的 CTR 都集中在 3% ～ 15%，内容之间 CTR 的差异都是在小数点后 1 ～ 3 位体现的。在运营中经常可以看到这样的数据：

短视频 A 的 CTR 是 7.83%，短视频 B 的 CTR 是 7.85%，短视频 C 的 CTR 是 7.8%。

它们三者之间的差异仅仅是 0.02% ～ 0.03%，实在太小了，对于日常运营中数据分析是个不小的挑战。为了能够更显著地放大差异，通常将要分析的短视频样本（例如当天的所有短视频）中 CTR 的最大值映射到 100，最小值映射到 0。这样所有短视频都分布在 0 ～ 100 的区间，然后再看数据的分布和集中程度。

数据归一化 / 标准化常用的算法包括极差法和 z-score 法。

a）极差法

极差法，是数据归一化最简单的算法，它将数据缩放到 0 ～ 1 之间。

极差法的算法，如表 1-6 所示。

表 1-6　极差法计算方法

转换后数据 = (转换前数据 – 最小值)/ 极差，其中极差 = 最大值 – 最小值

极差法不关心转换前的数据是正是负，其转换结果范围都会缩放到 0 ～ 1。

极差法是动态算法，若加入了新的数据，全部数据都需要重新计算，如表 1-7 所示。

表 1-7　源系统的数值化年龄

用户	A	B	C	D	E	F	G
年龄	11	32	19	64	49	37	26

首先确定最小值、最大值和极差：

最小值：11

最大值：64

极差：64−11 = 53

故用户年龄用极差法转换后如表 1-8 所示。

表 1-8　将年龄用极差法进行转换

用户	A	B	C	D	E	F	G
年龄	11	32	19	64	49	37	26
年龄（极差法）	0	0.39	0.15	1	0.72	0.49	0.28

b）z-score 法

z-score 法，即将某一列数值按比例缩放到统一的范围，其均值为 0，且方差为 1。

z-score 法的算法如表 1-9 和表 1-10 所示。

表 1-9　z-score 法计算方法

转换后数据 = (转换前数据 − 平均值)/ 标准差

表 1-10　源系统的数值化年龄

用户	A	B	C	D	E	F	G
年龄	11	32	19	64	49	37	26

首先确定平均值和标准差：

平均值：34

标准差：18.05547

故用户年龄用 z-score 法转换后如表 1-11 所示。

表 1-11　将年龄用 z-score 法进行转换

用户	A	B	C	D	E	F	G
年龄	11	32	19	64	49	37	26
年龄（z-score 法）	0	−0.11	−0.83	1.66	0.83	0.16	−0.44

注意：严格意义上的归一化和标准化略有差异，但在业务运营中可以适度忽略此差异。

（7）数据维度拆解与合并

由于源系统的数据规范多种多样，对相同业务含义数据存储各不相同，在进行数据预处理的时候就要进行维度的拆解和合并。

a）维度拆解

维度拆解，即将单一的维度拆分为多个子维度，以满足后续多维分析的需要。

最常见的场景是将地址信息进行拆分。例如将独立存储的地址信息拆分为国家、省份、城市、区域、街道等数据。

b）维度合并

维度合并，即将同一业务含义的维度进行合并和归一，减少数据的维度，降低数据存储要求，如图 1-6 所示。

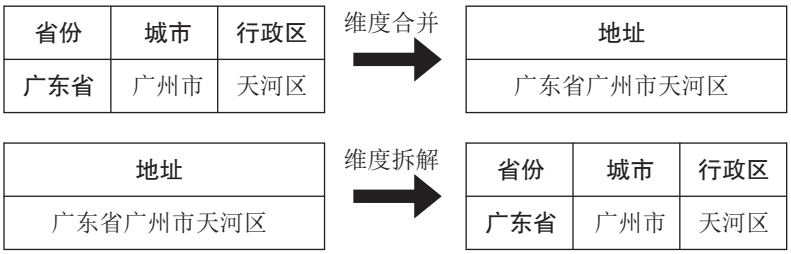

图 1-6　地址信息的维度拆解和合并

注意：在实际工程应用中，维度拆解是一项重要工作。高质量的维度拆解可以极大提升后续数据分析的速度和效果。正所谓，维度拆解做得好，数据分析没烦恼。

3. 加载（Load）

加载，即将清洗和转换后的数据加载到数据中间层或数据应用层，以供后续的数据分析使用。加载常见的有两种方式：全量加载和增量加载，通常与 Extract 抽取的策略一致。不同加载方式会影响最终报表平台上报表的创建

方式以及报表计算逻辑。

注意：关于 Load 加载的详细信息不是本书的主要内容，略去不讲。

1.2 用 Excel 完成常见数据预处理

前面讲了数据预处理的内容，本节开始就不能纸上谈兵了，我们用 Excel 来实现数据预处理中常见的操作。

1.2.1 文本数值化：文本数字转为数值型数字

在 Excel 中，数字通常有两种存储形式——数值型和文本型。

■ 数值型

这是数字在 Excel 中规范和正确的存储形式，其表现形式为默认居右对齐。以数值型存储的数字，可以完成各种数学运算，包括但不限于计数、求和、平均值、方差、标准差、最大值、最小值等。

■ 文本型

这是数字在 Excel 中的另一种存储形式，其表现形式为默认居左对齐，因为其本质依然是文本。以文本型存储的数字，只能完成计数，其他数学运算均不能完成，故而存在极大的限制。为了能够进行后续数据分析，必须将其转换为数值型数字。

提示：如何快速判断单元格中的数字是数值型还是文本型呢？一般情况下看它们默认的对齐方式，或观察单元格左上角有无绿色小三角。如有绿色小三角，则此单元格为文本型数值。

1. 用分列将文本型数字转换为数值型数字

操作步骤

第一步：选择需要转换的列，如图 1-7 所示。

第二步：单击“数据”选项卡，找到“分列”功能，如图 1-8 所示。

第三步：单击“分列”按钮，显示分列对话框，如图 1-9 所示。

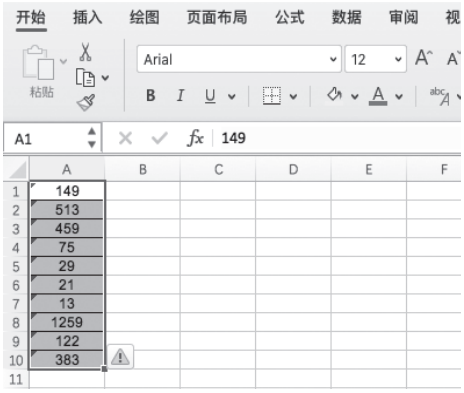

图 1-7　选择要转换的文本型数字

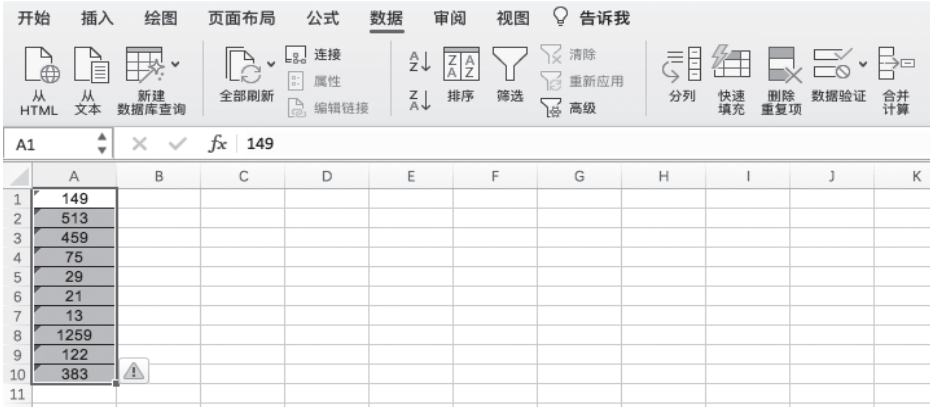

图 1-8　进入数据选项卡

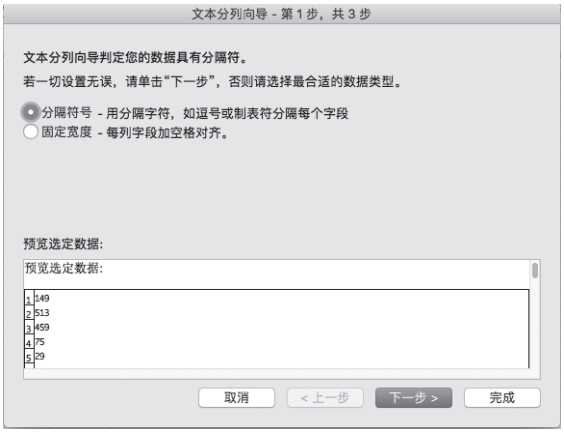

图 1-9　显示分列对话框

第四步：单击“完成”。

注意：不需要执行任何分列操作，只需要点击“完成”按钮即可，此时可以发现文本型数字已转换为数值型数字；亦可留意右下角，已显示平均值、计数和求和，证明此时已成为数值型数字。

2. 用公式将文本型数字转换为数值型数字

操作步骤

第一步：插入新列，作为转换辅助列，如图 1-10 所示。

第二步：在新列中使用 VALUE( ) 函数，如图 1-11 所示。

第三步：重复此操作，或双击单元格右下角的控制柄，如图 1-12 所示。

图 1-10　插入新列

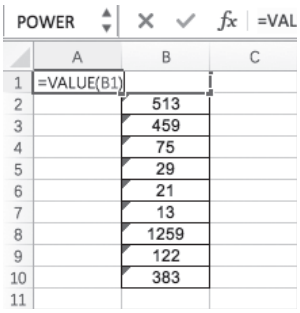

图 1-11　使用 VALUE( ) 函数

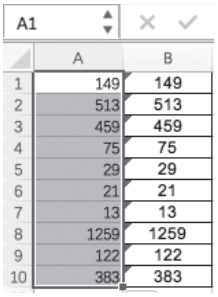

图 1-12　重复执行

汇总对比一下两种文本转为数字的方法，如表 1-12 所示。

表 1-12　两种文本转数字的方法对比

	上手难度	适合人群	缺　　点
分列	极易	所有人群	转换后的数字依然保留左对齐的特性，需要留意
公式	略难	有 Excel 函数经验的人群	需要熟悉 Excel 的 VBA 函数，需创建辅助列，且原始列数据不能删除

1.2.2　日期数值化：文本型日期转为日期型格式

在 Excel 中，日期通常有两种存储形式——日期型和文本型。

■ 日期型

这是日期在 Excel 中规范和正确的存储形式，其表现形式为默认居右对齐，并以特定的日期格式显示。以日期型存储的日期，可以完成各种操作，包括但不限于两个日期的差、提取星期、按日期的不同维度分组汇总和筛选等。

■ 文本型

这是日期在 Excel 中的另一种存储形式，其表现形式为默认居左对齐，因为其本质依然是文本。以文本型存储的日期，只能完成计数，其他均不能完成，故而存在极大的限制。为了能够进行后续数据分析，必须将其转换为日期型日期。

1. 用分列将文本型日期转换为日期型日期

操作步骤

第一步：选择需要转换的列，如图 1-13 所示。

图 1-13　选择要转换的文本型日期

第二步：单击“数据”选项卡，找到“分列”功能，如图 1-14 所示。

图 1-14　进入分列功能

第三步：单击“分列”按钮，显示分列对话框，如图 1-15 所示。

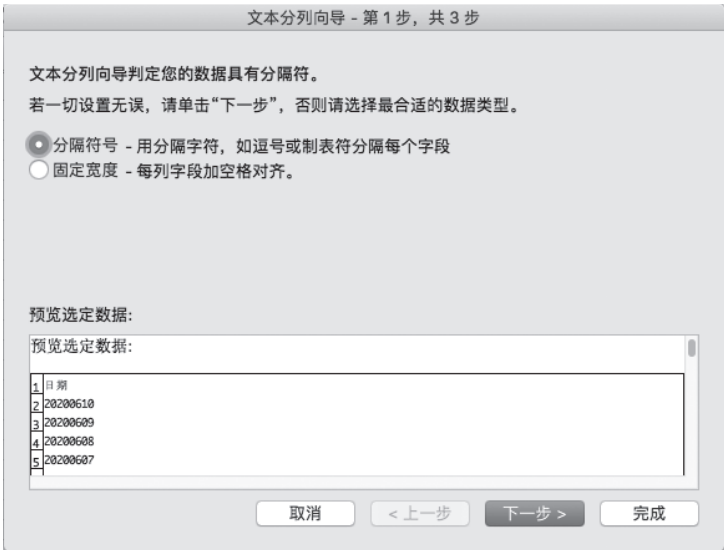

图 1-15　显示分列对话框

第四步：单击“完成”。

注意：不需要执行任何分列操作，只需要点击“完成”按钮即可，此时可以发现文本型日期已转换为日期型日期；亦可留意右下角，已显示平均值、计数和求和，证明此时已成为数值型数字。

特别提醒

此方法的前提是文本型日期依然遵循日期格式的样式来存储，目前支持如下格式的文本型日期通过分列操作转换为日期型：

- YYYY/MM/DD
- YYYY-MM-DD

2. 用公式将文本型日期转换为日期型日期

Excel 同样提供了将文本型日期转换为日期型日期的函数，即 DATEVALUE( ) 函数，使用方式为：DATEVALUE(“2009/01/01”)。两种转换方式的对比如表 1-13 所示。

表 1-13　两种文本转日期的方法对比

	上手难度	适合人群	缺　　点
分列	极易	所有人群	转换后的日期依然保留左对齐的特性，需要留意
公式	略难	有 Excel 函数经验的人群	需要熟悉 Excel 的 VBA 函数，需创建辅助列，且原始列数据不能删除

1.2.3　用分列实现维度拆分

在 Excel 中进行维度拆分是一件非常容易和轻松的工作，因为 Excel 为我们提供了“分列”功能。在很多源系统中地址类信息都是重要信息，包括用户联系地址、商品收货地址等，多数情况下是用单一字段进行存储，如表 1-14 所示。

表 1-14　地址信息

地　　址
广东省广州市天河区

此类型存储的地址信息对于数据分析是毫无意义的，因为它将大量有用信息融合进了同一字段，或者叫数据列，而所有分析工具的最小分析维度就是一个数据列。如果用这样的地址信息来分析可乐在不同城市的销量，你会发现无法下手，城市这个重要的分析目标恰恰被融合在一个夹杂很多无用数据的数据列中，数据分析工具无法处理数据列内的信息，就好像装着很多糖果的透明玻璃盒，你能看到却无法从中提炼出价值。

Excel 中分列的方式有两种：基于固定位置，基于分隔符。

1. 基于固定位置分列

基于固定位置分列，只需指定数据中需要分列的位置即可，支持同时指定多个分列位置，适合格式统一和固定的数据。

应用：提取身份证中的出生年月

第一步：单击“数据”选项卡，找到“分列”功能，如图 1-16 所示。

图 1-16　“分列”功能

第二步：单击“分列”按钮，显示分列对话框，如图 1-17 所示。

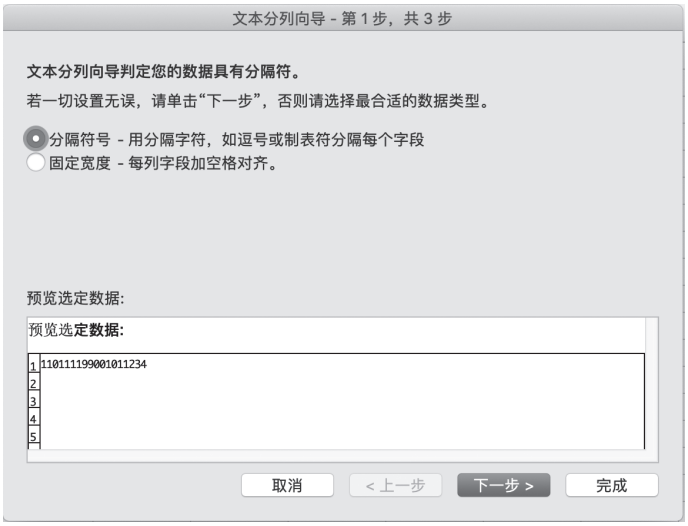

图 1-17　显示分列对话框

第三步：点击“固定宽度”并单击“下一步”按钮，结果如图 1-18 所示。

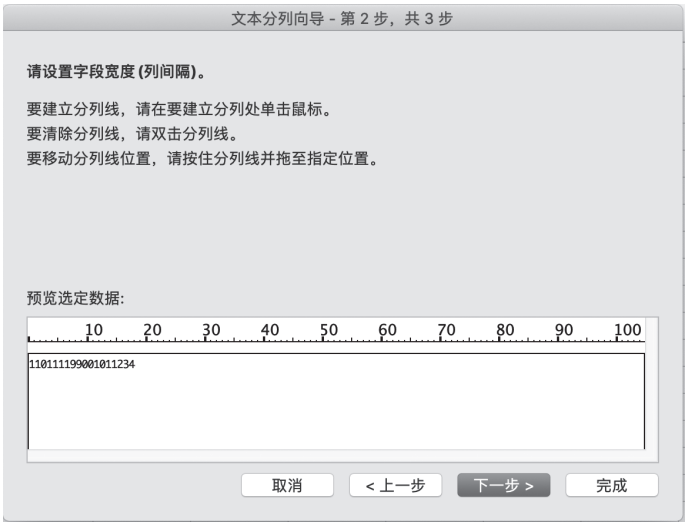

图 1-18　选择“固定宽度”

第四步：在标尺上单击创建分列线，拖动分列线到合适的位置并进入“下一步”，如图 1-19 所示。

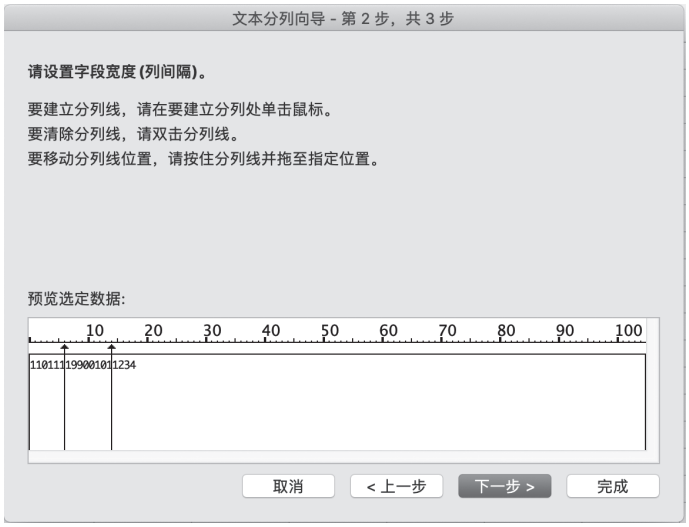

图 1-19　创建分列线

第五步：检查分列数据是否正确，无误后单击“完成”按钮，如图 1-20 所示。

图 1-20　完成分列

2. 基于分隔符分列

基于分隔符分列，需要指定数据中按照哪些字符进行拆分即可，每次分列仅支持一个分隔符，适合格式复杂的数据。

应用：提取地址信息中的省份

第一步：选择待拆分的列，如图 1-21 所示。

图 1-21　选择待拆分的列

第二步：单击“数据”选项卡，找到“分列”功能，如图 1-22 所示。

图 1-22　进入分列功能

第三步：单击“分列”按钮，显示分列对话框，如图 1-23 所示。

第四步：点击“分隔符号”并进入“下一步”，如图 1-24 所示。

第五步：在分隔符中点击“其他”输入“省”并进入“下一步”，如图 1-25 所示。

第六步：检查分列数据是否正确，无误后单击“完成”按钮。

提取地址中的信息一般只能用分隔符分列，因为省份名称长短不一，无法用固定宽度进行分列。

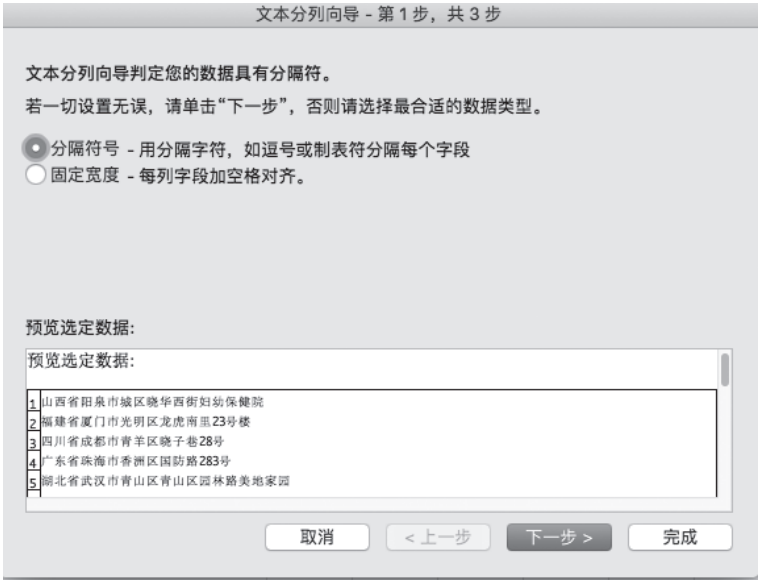

图 1-23　显示分列对话框

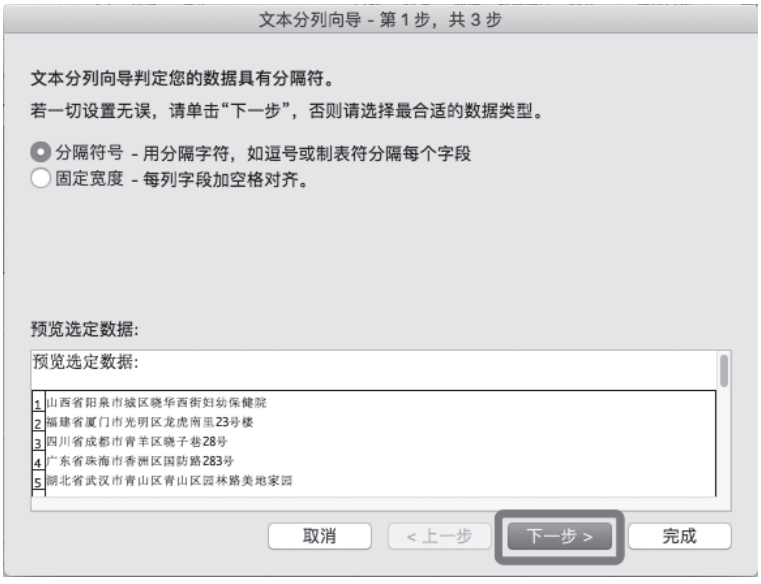

图 1-24　选择“分隔符号”

图 1-25　创建分隔符

注意：分列后的数据会覆盖相邻列，请提前插入空白列，否则相邻列的数据会被覆盖；分列后用于分列的字符会被删除，需检查分列后数据是否符合要求。再次强调，是否要做维度拆解，决定于拆解后的多个子维度是否为业务分析所需。

1.2.4　用“查找并删除重复行”处理重复值

在数据分析前必须仔细处理重复值。若源系统数据质量管控不严，极易产生大量重复数据。重复数据不仅占用大量存储空间，拖慢分析速度，甚至会直接导致分析结果出现错误或偏差。假如有 10 行性别数据，其男女比例是 6 ： 4，但其中有 4 个男性数据是重复的，故而真实的男女比例应为 3 ： 4。

Excel 为重复值处理提供了便捷的工具，即删除重复项。

应用：删除重复数据

第一步：选择要处理的数据区域，如图 1-26 所示。

第二步：单击“数据”选项卡，找到“删除重复项”功能，如图 1-27 所示。

第三步：单击“删除重复项”按钮，显示对话框，如图 1-28 所示。

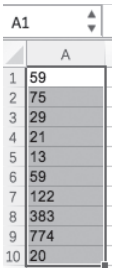

图 1-26　选择要处理的数据

第四步：勾选用以判断重复的列，如图 1-29 所示。

图 1-27　删除重复项功能

图 1-28　显示删除重复项对话框

图 1-29　选择要判断重复的列

第五步：单击“确定”按钮，如图 1-30 所示。

第六步：Excel 提示重复值数量，并自动保留唯一行，无误后单击“确定”按钮，如图 1-31 所示。

图 1-30　选择要判断重复的列

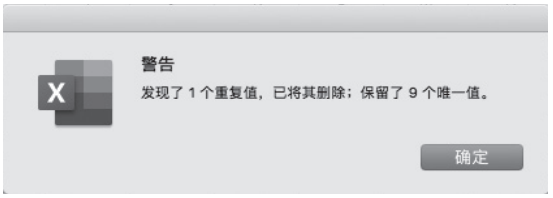

图 1-31　完成删除重复值

注意：勾选判断重复项的列时，建议尽可能选择完备。一般来说，重复列要求所有列的数据都一样，否则不能作为重复列进行删除处理。

1.3 本章小结和思考

1. 数据预处理的目的是什么？
2. 文本型数字如何转化为数值型数字？
3. 如何将一个字段拆分为多个字段？
4. 文本型日期和日期型日期的区别是什么？