第3章

数据可视化过程

人类视觉感知到心理认知的过程要经过信息的获取、分析归纳、存储、概念、提取、使用等一系列加工阶段。尽管不同领域的数据可视化面向不同数据，面临不同的挑战，但可视化的基本步骤和流程是相同的。本章将学习从社会自然现象数据中提取信息、知识和灵感的可视化基本流程。

3.1 数据可视化流程

可视化不是一个单独的算法，而是一个流程。除了视觉映射外，也需要设计并实现其他关键环节，如前端的数据采集、处理和后端的用户交互。这些环节是解决实际问题必不可少的步骤，且直接影响可视化效果。作为可视化设计者，解析可视化流程有助于把问题化整为零，降低设计的复杂度。作为可视化开发者，解析可视化流程有助于软件开发模块化、提高开发效率、缩小问题范围、重复利用代码，有助于设计工具库、编程界面和软件模块。

数据可视化是一个流程，有点像流水线，但这些流水线之间是可以相互作用的、双向的。可视化流程以数据流为主线，主要包括数据采集、数据处理和变换、可视化映射、用户感知模块。图 3-1 所示的是一个数据可视化流程。

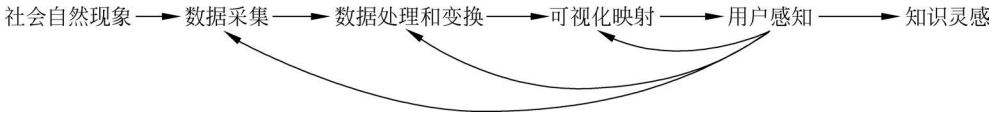


图 3-1　数据可视化流程

(1) 数据采集。数据的采集直接决定了数据的格式、维度、尺寸、分辨率、精确度等重要性质，在很大程度上决定了可视化结果的质量。

(2) 数据处理和变换。数据处理和变换是可视化的前期处理。一方面，原始数据不可

避免地含有噪声和误差；另一方面，数据的模式和特征往往被隐藏。而可视化需要将难以理解的原始数据变换成用户可以理解的模式和特征并显示出来。这个过程包括去除数据噪声、数据清洗、提取特征等，为之后的可视化映射做准备。

(3) 可视化映射。可视化映射是整个可视化流程的核心，它将数据的数值、空间位置、不同位置数据间的联系等，映射到不同的视觉通道，如标记、位置、形状、大小和颜色等。这种映射的最终目的是让用户通过可视化，洞察数据和数据背后隐含的现象和规律。因此可视化映射的设计不是一个孤立的过程，而是与数据、感知、人机交互等方面相互依托，共同实现的。

(4) 用户感知。数据可视化和其他数据分析处理办法的最大不同是用户的关键作用。用户借助数据可视化结果感受数据的不同，从中提取信息、知识和灵感。可视化映射后的结果只有通过用户感知才能转换成知识和灵感。用户感知可以在任何时期反作用于数据的采集、处理变换以及映射过程中，如图 3-1 所示。

数据可视化可用于从数据中探索新的假设，也可证实相关假设与数据是否吻合，还可以帮助专家向公众展示数据中的信息。用户的作用除被动感知外，还包括与可视化其他模块的交互。交互在可视化辅助分析决策中发挥了重要作用。有关人机交互的探索已经持续了很长时间，但智能、适用于海量数据可视化的交互技术，如任务导向的、基于假设的方法还是一个未解难题。

上面的可视化流程虽然简单，但也要注意以下两点。

(1) 上述过程都是基于数据背后的自然现象或者社会现象，而不是数据本身。

(2) 图 3-1 中的各个模块的联系并不是顺序的线性的联系，它们之间的联系更多的是非线性的，任意两个模块之间都可能存在联系。

3.2 数据处理和数据变换

在可视化流程中，原始数据经过处理和变换后得到清洁、简化、结构清晰的数据，并输出到可视化映射模块中。数据处理和变换直接影响到可视化映射的设计，对可视化的最终结果也有重要的影响。

当今现实世界的数据库极易受噪声、缺失值和不一致数据的侵扰，有大量数据预处理技术。数据清理可以清除数据中的噪声，纠正不一致；数据集成将数据由多个数据源合并成一致的数据存储，如数据仓库；数据归约可以通过如聚集、删除冗余特征或聚类来降低数据的规模；数据变换(如规范化)可以用来把数据压缩到较小的区间，如[0.0,1.0]，这可以提高涉及距离度量的挖掘算法的精确率和效率。这些技术不是相互排斥的，可以一起使用。例如，数据清理可能涉及纠正错误数据的变换，如通过把一个数据字段的所有项都变换成公共格式进行数据清理。

数据如果能满足其应用要求，那么它是高质量的。数据质量涉及许多因素，包括准确性、完整性、一致性、时效性、可信性和可解释性。

数据处理和数据变换主要步骤：数据清理、数据集成、数据变换与数据离散化以及数据配准。

3.2.1 数据清理

现实世界的数据一般是不完整的、有噪声的和不一致的。数据清理试图填充缺失的值，光滑噪声和识别或删除离群点，并纠正数据中的不一致来清理数据。

1. 缺失值

假设分析某公司 AllElectronics 的销售和客户数据，发现许多记录的一些属性（如客户的 income）没有记录值。那么应怎样处理该属性缺失的值呢？可用的处理方法如下。

（1）删除记录。删除属性缺少的记录简单直接，代价和资源较少，并且易于实现，然而直接删除记录会浪费该记录中被正确记录的属性。当属性缺失值的记录百分比很大时，它的性能特别差。

（2）人工填写缺失值。一般地说，该方法很费时，并且当数据集很大、缺少很多值时，该方法可能行不通。

（3）使用一个全局常量填充缺失值。将缺失的属性值用同一个常量（如 Unknown 或-）替换。如果缺失的值都用 Unknown 替换，则挖掘程序可能误以为它们形成了一个有趣的概念，因为它们都具有相同的值——Unknown。因此，尽管该方法简单，但是并不十分可靠。

（4）使用属性的中心度量（如均值或中位数）填充缺失值。对于正常的（对称的）数据分布而言，可以使用均值，而倾斜数据分布应该使用中位数。例如，假定 AllElectronics 的客户的平均收入为 18 000 美元，则使用该值替换 income 中的缺失值。

（5）使用与属性缺失的记录属同一类的所有样本的属性均值或中位数。例如，如果将客户按 credit_risk 分类，则用具有相同信用风险的客户的平均收入替换 income 中的缺失值。如果给定类的数据分布是倾斜的，则中位数是更好的选择。

（6）使用最可能的值填充缺失值。可以用回归、贝叶斯形式化方法的推理工具或决策树归纳确定。例如，利用数据集里其他客户的属性，可以构造一棵判定树来预测 income 的缺失值。

方法（3）～方法（6）使数据有偏差，填入的值可能不正确。然而方法（6）是最流行的策略。与其他方法相比，它使用已有记录（数据）的其他部分信息来推测缺失值。在估计 income 的缺失值时，通过考虑其他属性的值，有更大的机会保持 income 和其他属性之间的联系。

在某些情况下，缺失值并不意味着有错误；在理想情况下，每个属性都应当有一个或多个关于空值条件的规则。这些规则可以说明是否允许空值，并且/或者说明这样的空值应当如何处理或转换。

2. 噪声数据与离群点

噪声是被测量变量的随机误差（一般指错误的数据）。离群点是数据集中包含的一些数据对象，它们与数据的一般行为或模型不一致（正常值，但偏离大多数数据）。例如，在图 3-2 中出现了负年龄（噪声数据），以及 85～90 岁的用户（离群点）。

给定一个数值属性，可以采用下面的数据光滑技术“光滑”数据，去掉噪声。

（1）分箱。

分箱方法通过考查数据的“近邻”（即周围的值）来光滑有序数据值。这些有序的值被分

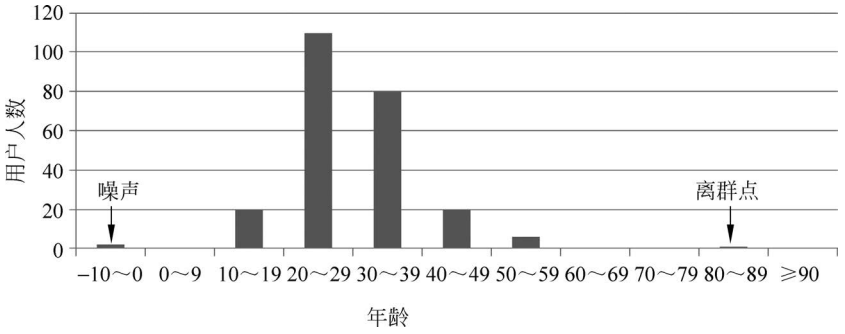


图 3-2 系统用户年龄的分析

布到一些桶或箱中。由于分箱方法考查近邻的值,因此对它进行局部光滑。

① 用箱均值光滑。箱中每一个值被箱中的平均值替换。

② 用箱边界光滑。箱中的最大值和最小值同样被视为边界。箱中的每一个值被最近的边界值替换。

③ 用箱中位数光滑。箱中的每一个值被箱中的中位数替换。

如图 3-3 所示,数据首先排序并被划分到大小为 3 的等深的箱中。对于用箱均值光滑,箱中每一个值都被替换为箱中的均值。类似地,可以使用箱边界光滑或者箱中位数光滑等。

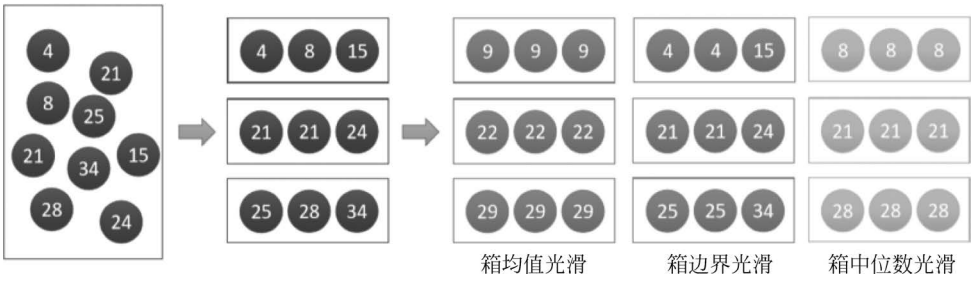


图 3-3 数据光滑的分箱方法

上面分箱的方法采用等深分箱(每个“桶”的样本个数相同),也可以是等宽分箱(其中每个箱值的区间范围相同)。一般而言,宽度越大,光滑效果越明显。分箱也可以作为一种离散化技术使用。

(2) 回归。

回归(Regression)用一个函数拟合数据来“光滑”数据。线性回归涉及找出拟合两个属性(或变量)的“最佳”直线,使得一个属性能够预测另一个。图 3-4 即对数据进行线性回归拟合。图 3-4 中已知有 10 个点,此时获得信息将在横坐标 7 的位置出现一个新的点,却不知道纵坐标,请预测最有可能的纵坐标值。这是典型的预测问题,可以通过回归来实现。预测结果如图 3-4 所示,预测点采用菱形标出。

多线性回归是线性回归的扩展,它涉及多于两个属性,并且数据拟合到一个多维面。使用回归,找出适合数据的拟合函数,能够帮助消除噪声。

离群点分析可以通过如聚类来检测离群点。聚类将类似的值组织成群或“簇”。直观地落在簇集合之外的值被视为离群点。

图 3-5 所示为聚类出 3 个数据簇。可以将离群点看作落在簇集合之外的值来检测。

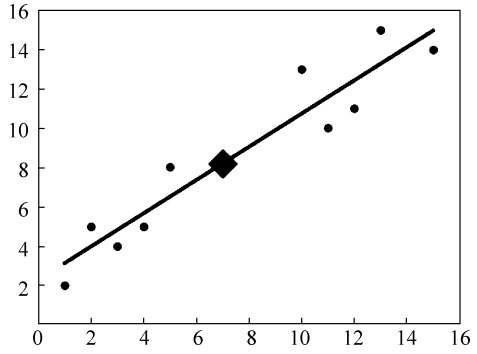


图 3-4　线性回归拟合

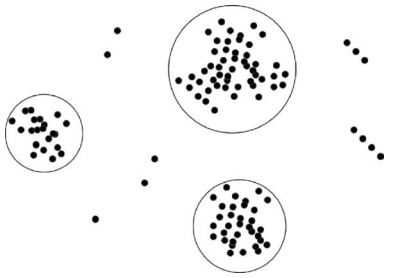


图 3-5　聚类出 3 个数据簇

许多数据光滑的方法也用于数据离散化(一种数据变换方式)和数据归约。例如,上面介绍的分箱技术减少了每个属性不同值的数量。对于基于逻辑的数据挖掘方法(决策树归纳),数据离散化充当了一种形式的数据归约。概念分层是一种数据离散化形式,也可以用于数据平滑。

3. 不一致数据

对于有些事务,所记录的数据可能存在不一致。有些数据不一致可以根据其他材料上的信息人工地加以更正。例如,数据输入时的错误可以使用纸上的记录加以更正,也可以用纠正不一致数据的程序工具来检测违反限制的数据。例如,知道属性间的函数依赖,可以查找违反函数依赖的值。

3.2.2　数据集成

上述数据清理方法一般应用于同一数据源的不同数据记录上。在实际应用中,经常会遇到来自不同数据源的同类数据,且在用于分析之前需要进行合并操作。实施这种合并操作的步骤称为数据集成。有效的数据集成过程有助于减少合并后的数据冲突,降低数据冗余程度等。

数据集成需要解决的问题如下。

1. 属性匹配

对于来自不同数据源的记录,需要判定记录中是否存在重复记录。而首先需要做的是确定不同数据源中数据属性间的对应关系。例如,从不同销售商收集的销售记录可能对用户 id 的表达有多种形式(销售商 A 使用 cus_id,数据类型为字符串;销售商 B 使用 customer_id_number,数据类型为整数),在进行销售记录集成之前,需要先对不同的表达方式进行识别和对应。

2. 冗余去除

数据集成后产生的冗余包括两方面:数据记录的冗余,例如,Google 街景车在拍摄街景照片时,不同的街景车可能有路线上的重复,这些重复路线上的照片数据在进行集成时便会造成数据冗余(同一段街区被不同车辆拍摄);因数据属性间的推导关系而造成数据属性冗余,例如,调查问卷的统计数据中,来自地区 A 的问卷统计结果注明了总人数和男性受调

查者人数,而来自地区B的问卷统计结果注明了总人数和女性受调查者人数,当对两个地区的问卷统计数据进了集成时,需要保留“总人数”这一数据属性,而“男性受调查者人数”和“女性受调查者人数”这两个属性保留一个即可,因为两者中任一属性可由“总人数”与另一属性推出,从而避免了在集成过程中由于保留所有不同数据属性(即使仅出现在部分数据源中)而造成的属性冗余。

3. 数据冲突检测与处理

来自不同数据源的数据记录在集成时因某种属性或约束上的冲突,导致集成过程无法进行。例如,当来自两个不同国家的销售商使用的交易货币不同时,无法将两份交易记录直接集成(涉及货币单位不同这一属性冲突)。

数据挖掘和数据可视化经常需要数据集成——合并来自多个数据存储的数据。谨慎集成有助于减少结果数据集的冗余和不一致。这有助于提高其后挖掘和数据可视化过程的准确性和速度。

3.2.3 数据变换与数据离散化

在数据处理阶段,数据被变换或统一,使得数据可视化分析更有效,挖掘的模式可能更容易理解。数据离散化是一种数据变换形式。

1. 数据变换策略概述

数据变换策略包括以下6种。

(1) 光滑。去掉数据中的噪声。这种技术包括分箱、聚类和回归。

(2) 属性构造(或特征构造)。可以由给定的属性构造新的属性并添加到属性集中,以帮助挖掘过程。

(3) 聚集。对数据进行汇总和聚集。例如,可以聚集日销售数据,计算月和年销售量。通常这一步用来为多个抽象层的数据分析构造数据立方体。

(4) 规范化。把属性数据按比例缩放,使之落入一个特定的小区间,如[-1.0,1.0]或[0.0,1.0]。

(5) 离散化。数值属性(如年龄)的原始值用区间标签(如[0,10],[11~20]等)或概念标签(如youth、adult、senior)替换。这些标签可以递归地组织成更高层概念,导致数值属性的概念分层。

(6) 由标称数据产生概念分层。属性如street,可以泛化到较高的概念层,如city或country。

2. 通过规范化变换数据

规范化数据可赋予所有属性相等的权重。有许多数据规范化的方法,常用的是最小—最大规范化、z-score规范化和小数定标规范化。

下面令A是数值属性,具有n个值$v_1,v_2,\cdots,v_n$,采用这三种规范化方法变换数据。

(1) 最小—最大规范化是对原始数据进行线性变换。假定$\max_A$和$\min_A$分别为属性A的最大值和最小值。最小—最大规范化通过计算公式:

$$v_i'=\frac{v_i-\min_A}{\max_A-\min_A}(\text{new\_max}_A-\text{new\_min}_A)+\text{new\_min}_A$$

把 A 的值 v_i 映射到区间[new_min_A，new_max_A]中的 v_i'。最小—最大规范化保持原始数据值之间的联系。如果属性 A 的实际测试值落在 A 的原数据值域[$\min_A$，$\max_A$]之外，则该方法将面临“越界”错误。

(2) 在 z-score 规范化(或零—均值规范化)中，基于 A 的平均值和标准差规范化。A 的值 v_i 被规范化为 v_i'，由下式计算：

$$v_i'=\frac{v_i-\text{avg}_A}{\delta_A}$$

式中，avg_A 和 δ_A 分别为属性 A 的平均值和标准差。当属性 A 的实际最大值和最小值未知，或离群点左右了最小—最大规范化时，该方法是有用的。

(3) 小数定标规范化通过移动属性 A 的值的小数点位置进行规范化。小数点的移动位数依赖于 A 的最大绝对值。A 的值 v_i 被规范化为 v_i'，由下式计算：

$$v_i'=\frac{v_i}{10^j}$$

式中，j 为使得 $\max(|v_i'|)<1$ 的最小整数。

3. 通过分箱离散化

分箱是一种基于指定的箱个数的自顶向下的分裂技术。前面光滑噪声时已经介绍过。

分箱并不使用分类信息，因此是一种非监督的离散化技术。它对用户指定的箱个数很敏感，也容易受离群点的影响。

4. 通过直方图分析离散化

像分箱一样，直方图分析也是一种非监督离散化技术，因为它也不使用分类信息。直方图把属性 A 的值划分成不相交的区间，称作桶或箱。桶安放在水平轴上，而桶的高度(和面积)是该桶所代表值的出现频率。通常，桶表示给定属性的一个连续区间。

可以使用各种划分规则定义直方图。例如，在图 3-6 的直方图中，将值分成相等分区或区间(如属性“价格”，其中每个桶宽度为 10 美元)。

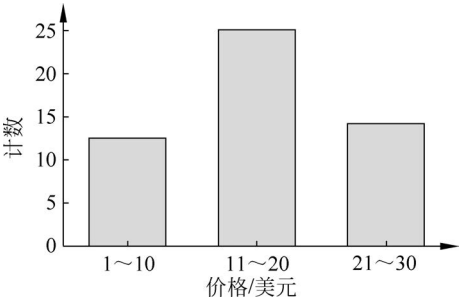


图 3-6　直方图

5. 通过聚类、决策树离散化

聚类分析是一种流行的离散化方法。通过将属性 A 的值划分成簇或组，聚类算法可以用来离散化数值属性 A。聚类考虑 A 的分布以及数据点的邻近性，因此可以产生高质量的离散化结果。

为分类生成决策树的技术可以用来离散化。这类技术使用自顶向下划分方法。离散化的决策树方法是监督的，因为它使用分类标号。其主要思想是选择划分点，使得一个给定的结果分区包含尽可能多的同类记录。

6. 标称数据的概念分层产生

概念分层可以用来把数据变换到多个粒度值。例如，由用户或专家在模式级显式地说明属性的部分序或全序，可以很容易地定义概念分层。例如，关系数据库或数据仓库的维

location 可能包含如下一组属性：street、city、province_or_state 和 country。可以在模式级说明一个全序，如 street < city < province_or_state < country，来定义分层结构。

使用概念分层变换数据使得较高层的知识模式可以被发现。

3.2.4 数据配准

数据可视化往往需要在同一空间中显示不同时间、不同角度、不同仪器或模拟算法产生的数据。例如，医生在观察病人的医学图像时，会比较当前的图像和该病人以前扫描的图像或健康人的图像，观察其异同。气象专家在观察气象数据时，会比较模拟算法产生的结果、气象台观测数据以及卫星图片等。这种不同数据之间的比较需要在同一空间中配准。图 3-7 所示的是数据配准过程。不同尺寸、方向的数据通过配准统一取目标数据的尺寸和方向。配准后更便于数据比较和发现细微的不同点。

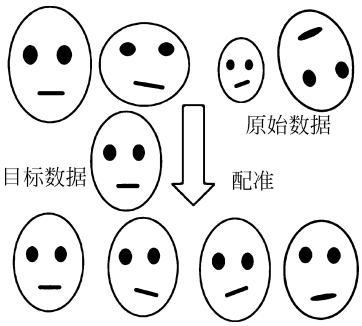


图 3-7 数据配准过程

数据配准的方法很多，在空间数据场分析和可视化中应用广泛，如医学影像处理。实现两个空间数据场的配准，大多需要计算两个数据之间的相似度，并通过对其中一个数据场的位移和变形来提高两者的相似度，以达到数据配准的目的。按计算相似度的方式，可以将数据配准分为基于像素强度的方法和基于特征的方法。基于像素强度的方法，用数据场采样点的强度的分布计算两个数据的相似度；而基于特征的方法，则用数据场中的特征，如点、线、等值线检测两者的相似度。

在可视化中还经常用到数据转换函数，如将数据的取值映射到显示像素的强度范围内(规范化)；对数据进行统计，如计算其平均值和方差；或变换数据的分布(如将指数分布的数据用对数函数转换为直线分布)等。当数据经过这些变换后，需要告知用户变换的函数和目的，以帮助用户分析可视化，避免解读上的偏差。

3.3 可视化映射

简单来讲，人类视觉的特点如下。

(1) 对亮度、运动、差异更敏感，对红色相对于其他颜色更敏感。

(2) 对于具备某些特点的视觉元素具备很强的识别能力，如空间距离较近的点往往被认为具有某些共同的特点。

(3) 对眼球中心正面物体的分辨率更高，这是由于人类晶状体中心区域锥体细胞分布最为密集。

(4) 人在观察事物时习惯于将具有某种方向上的趋势的物体视为连续物体。

(5) 人习惯于使用经验去感知事物整体，而忽略局部信息。

根据人类视觉特点，将数据信息映射成可视化元素，这里引入一个概念——可视化映射(或称可视化编码，Visual Encoding)。可视化映射是数据可视化的核心步骤，指将数据信息映射成可视化元素。映射结果通常具有表达直观、易于理解和记忆等特性。数据对象由属

性描述。例如，在学生成绩数据中，学生数据对象由学号、姓名、成绩等属性组成，“学号”属性取值为数字串，“姓名”属性取值为字符串，“成绩”属性取值为数字。属性和它的值对应可视化元素分别是图形标记和视觉通道。

3.3.1　图形标记和视觉通道

可视化映射(可视化编码)是信息可视化的核心内容。数据通常有属性和它的值，因此可视化编码类似地由图形标记和视觉通道两方面组成。图形标记通常是一些几何图形元素，如点、线、面等，如图 3-8 所示。视觉通道用于控制标记的视觉特征。

1. 图形标记维度

根据图形标记代表的数据维度来划分，图形标记分为如下 4 种。

(1) 零维。点是常见的零维图形标记，点仅有位置信息。

(2) 一维。常见的一维图形标记是直线。

(3) 二维。常见的二维标记是二维平面。

(4) 三维。常见的立方体、圆柱体都是三维的图形标记。

图形标记可以代表的数据维度如图 3-8 所示。

2. 图形标记自由度

前面我们介绍过坐标系，坐标系代表了图形所在的空间维度；而图形空间的自由度是在不改变图形性质的基础下可以自由扩展的维度，即自由度＝空间维度－图形标记的维度。那么：

(1) 点在二维空间内的自由度是 2，即可以沿 x 轴、y 轴方向进行扩展。

(2) 线在二维空间内的自由度是 1，即线仅能增加宽度，而无法增加长度。

(3) 面在二维空间内的自由度是 0，以一个多边形为示例，在不改变代表多边形数据的前提下，我们无法增加多边形的宽度或者高度。

(4) 面在三维空间的自由度是 1，可以更改面的厚度。

图形标记可以代表的数据自由度如图 3-9 所示。

图 3-8　图形标记的数据维度

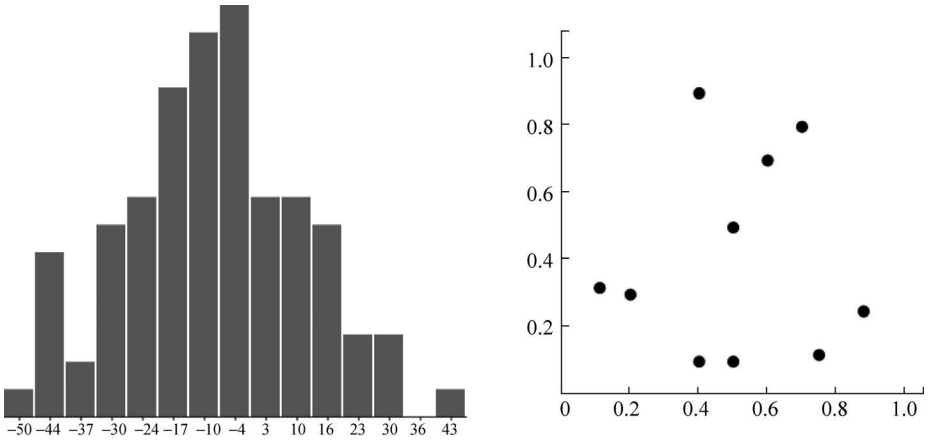
图 3-9　图形标记的数据自由度

3. 可视化表达常用的视觉通道

第 2 章已经介绍了可视化视觉通道。视觉通道用于控制标记的视觉特征，通常可用的视觉通道包括位置、大小、形状、方向、色调、饱和度、亮度等(见第 2 章的图 2-14)。例如，对于柱状图[见图 3-10(a)]而言，图形标记就是矩形，视觉通道就是矩形的颜色、高度或宽度等。对于散点图[见图 3-10(b)]而言，图形标记就是点，视觉通道就是竖直位置和水平位置，这样达到数据编码的目的。图形标记的自由度与数据能够映射到图形的视觉通道数量相关。

高效的可视化可以使用户在较短的时间内获取原始数据更多、更完整的信息，而其设计的关键因素是视觉通道的合理运用。

数据可视化的设计目标和制作原则在于信、达、雅，即一要精准展现数据的差异、趋势、



(a) (b)

图 3-10 柱状图和散点图

规律；二要准确传递核心思想；三要简洁美观，不携带冗余信息。结合人的视觉特点，很容易总结出好的数据可视化作品的基本特征。

(1) 让用户的视线聚焦在可视化结果中最重要的部分。

(2) 对于有对比需求的数据，使用亮度、大小、形状来进行编码更佳。

(3) 使用尽量少的视觉通道数据编码，避免干扰信息。

3.3.2 可视化编码的选择

图形标记的选择通常基于人们对于事物理解的直觉。然而，不同的视觉通道在表达信息的作用和能力上可能具有截然不同的特性。可视化设计人员必须了解和掌握每个视觉通道的特性以及它们可能存在的相互影响。例如，可视化设计中应该优选哪些视觉通道？具体有多少不同的视觉通道可供使用？某个视觉通道能编码什么信息，能包含多少信息量？视觉通道表达信息能力有什么区别？哪些视觉通道不相关，而哪些又相互影响？只有熟知视觉通道的特点，才能设计出有效解释数据信息的可视化。图 3-11 所示的是视觉通道在数值型数据可视化编码的优先级。

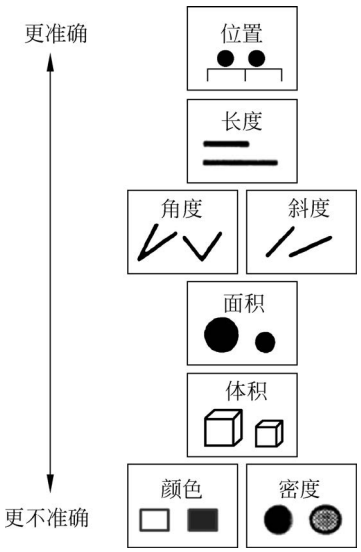


图 3-11 视觉通道在数值型数据可视化编码的优先级

显然，可视化的对象不仅是数值型数据，也包含非数值型数据。图 3-11 的排序对数值型可视化有指导意义，但对非数值型数据并不通用。例如，颜色对区分不同种类数据非常有效，但它排在图 3-11 的底层。

图 3-12 显示视觉通道的可视化元素对数值型数据、序列型数据和类别型数据的有效性排序。不同视觉通道元素在这三种数据中的排序不一样，又有一定的联系。例如，标记的位置是最准确反映各种类型数据的可视化元素。颜色对数值型数据的映射效果不佳，却能很好地反映类别型数据甚至序列型数据。而长度、角度和方向

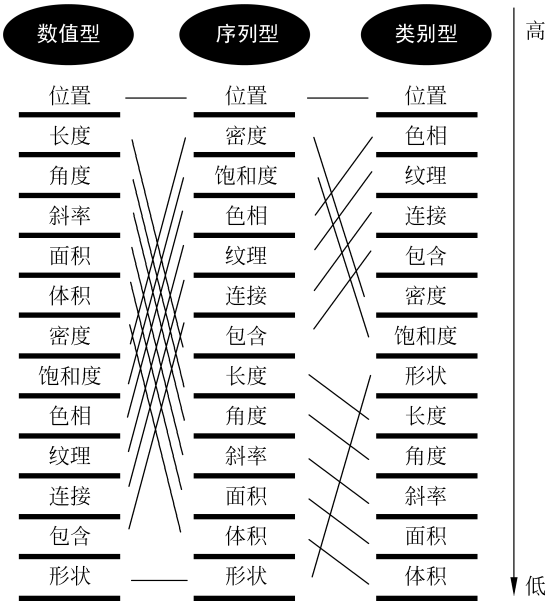


图 3-12　视觉通道在不同数据可视化编码的优先级

等元素对数值型数据有很好的效果，却不能很好地反映序列型数据和类别型数据。

从图 3-12 可以看出，数据可视化中常用的视觉编码通道，针对同种数据类型，采用不同的视觉通道带来的主观认知差异很大。数值型适合用能够量化的视觉通道表示，如坐标、长度等，使用颜色表示的效果就大打折扣，且容易引起歧义；类似地，序列型适合用区分度明显的视觉通道表示；类别型适合用易于分组的视觉通道表示。

需要指出的是，图 3-12 蕴含的理念可以应对绝大多数应用场景下可视化图形的设计“套路”，但数据可视化作为视觉设计的本质决定了“山无常势，水无常形”，任何可视化效果都拒绝生搬硬套，更不要说数据可视化的应用还要受到业务、场景和受众的影响。

3.3.3　源于统计图表的可视化

统计图表是使用最早的可视化图形，在数百年的发展过程中，逐渐形成了基本“套路”，符合人类感知和认知，进而被广泛接受。

常见于各种统计分析报告的有柱状图、折线图、饼图、散点图、气泡图、雷达图。在可视化设计中，我们将常见的图形标记定义成图表类型。下面了解一下最常用的图表类型。

1. 柱状图

柱状图(Bar Chart)是最常见的图表之一，也最容易解读。它的适用场合是二维数据集(每个数据点包括两个值 x 和 y)，但只有一个维度需要比较。如图 3-13 所示，月销售额就是二维数据，“月份”和“销售额”就是它的两个维度，但只需要比较“销售额”这一个维度。

柱状图利用柱子的高度，反映数据的差异。肉眼对高度差异很敏感，辨识效果非常好。柱状图的局限在于只适用中小规模的数据集。

2. 折线图

折线图(Line Chart)是用直线段将各数据点连接起来而组成的图形，以折线方式显示

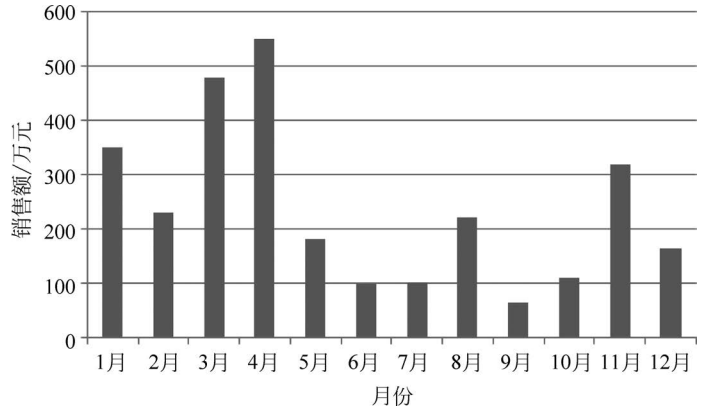


图 3-13 月销售额柱状图

数据的变化趋势和对比关系。折线图可以显示随时间而变化的连续数据,因此非常适用于显示在相等时间间隔下数据的趋势。

折线图适合二维的大数据集,尤其适合研究趋势的场合。它还适合多个二维数据集的比较。图 3-14 所示的是一个二维数据集(月销售额)的折线图。

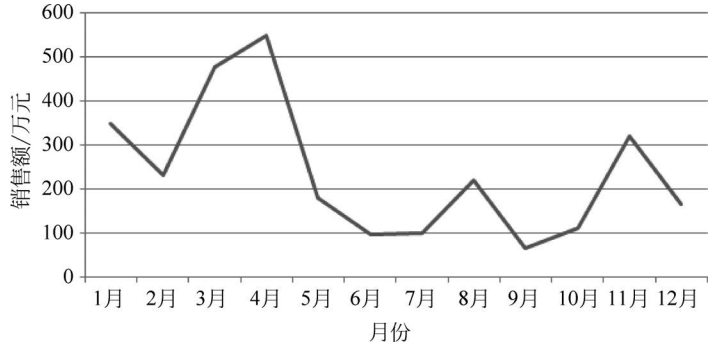


图 3-14 月销售额折线图

3. 饼图

饼图(Pie Chart)是用扇形面积,也就是圆心角的度数来表示数量。饼图可以根据圆中各个扇形面积的大小来判断某一部分在总体中所占比例的多少。饼图是一种应该避免使用的图表,因为肉眼对面积大小不敏感。

图 3-15(a)中饼图的五个色块的面积排序不容易看出来;若换成图 3-15(b)中的柱状图,就容易多了。一般情况下,总是用柱状图替代饼图。但有一个例外,就是反映某部分占整体的比重情况,如贫穷人口占总人口的百分比。

4. 散点图

散点图(Scatter Chart)表示因变量随自变量而变化的大致趋势,据此可以选择合适的函数对数据点进行拟合。散点图通常用于显示和比较数值,如科学数据、统计数据和工程数据。当不考虑时间的情况而比较大量数据点时,散点图就是最好的选择。散点图中包含的数据越多,比较的效果就越好。在默认情况下,散点图以圆点显示数据点。如果在散点图中有多个序列,可考虑将每个点的标记形状更改为方形、三角形、菱形或其他形状。散点图适

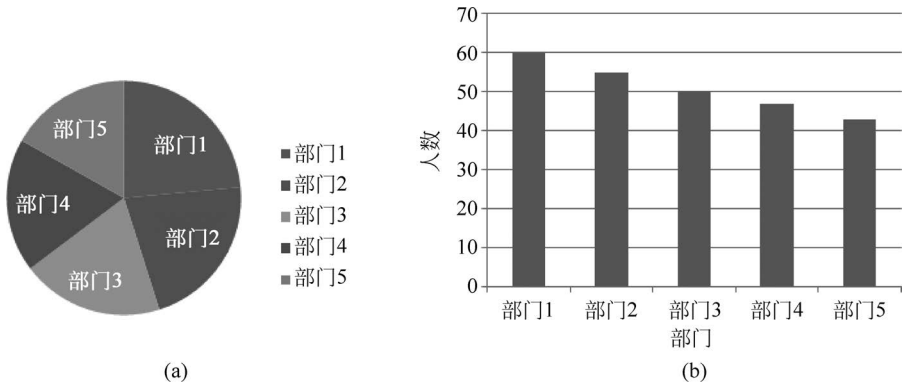


图 3-15　饼图和柱状图对比

用于两维比较。

图 3-16(a)所示的是普通的散点图，数据点的分布展示了不同年龄段的月均网购金额，从图表中可以分析出，月均网购金额较高的人群主要集中 30 岁左右。但是，对比图 3-16(b)后发现，在连续的年龄段上，图 3-16(a)中数据较密的点不容易区分，而图 3-16(b)中将所有数据点通过年龄的增加联系起来，不但表示了数据本身的分布情况，还表示了数据的连续性。用带平滑线(函数拟合)和数据标记的散点图来表示这样的数据比普通的散点效果更好。

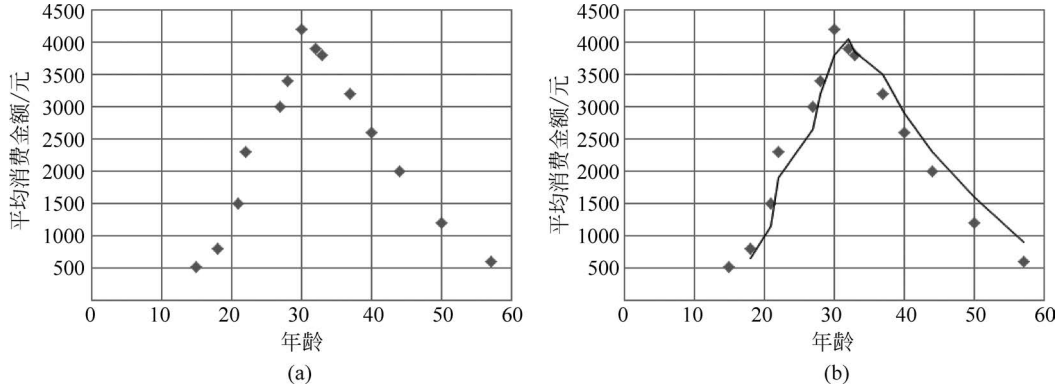


图 3-16　散点图

5. 气泡图

气泡图(Bubble Chart)是散点图的一种变体，通过每个点的面积大小，反映第三维。图 3-17 是气泡图，显示“卡特里娜”飓风的路径，三个维度分别为经度、纬度、强度。点的面积越大，就代表强度越大。因为用户不善于判断面积大小，所以气泡图只适用不要求精确辨识第三维的场合。

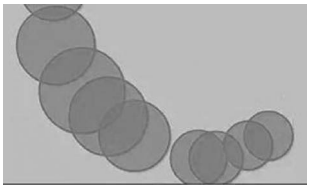


图 3-17　气泡图

如果为气泡加上不同颜色(或文字标签)，气泡图就可用来表达四维数据。如通过颜色，表示每个点的风力等级。

6. 雷达图

雷达图(Radar Chart)将多个维度的数据量映射到坐标轴上。这些坐标轴起始于同一个圆心点，通常结束于圆周边缘，将同一组的点使用线连接起来就成了雷达图，如图 3-18 所

示。雷达图适用于多维数据(四维以上),且每个维度必须可以排序。但是,它有一个局限,就是数据点最多为 6 个,否则无法辨别,因此适用场合有限。需要注意的是,如果用户不熟悉雷达图,在解读时就有困难。因此使用雷达图时应尽量标注说明,以减轻解读负担。

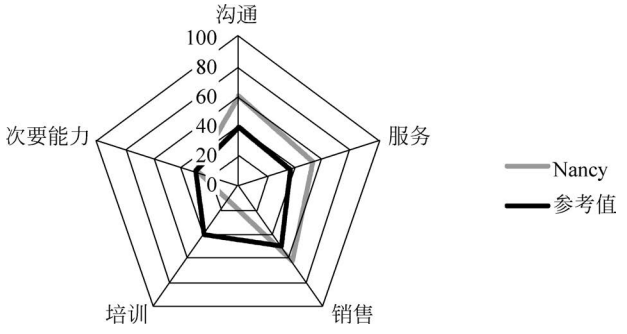


图 3-18 雷达图

7. 直方图

直方图(Histogram)又称质量分布图,是一种统计报告图,也是数据属性频率的统计工具。直方图由一系列高度不等的纵向条纹或线段表示数据分布的情况,一般用横轴表示数据类型,纵轴表示分布情况。例如,某次考试成绩分布如表 3-1 所示,对应直方图如图 3-19 所示。

表 3-1 某次考试成绩分布

分数段	0～29	30～39	40～49	50～59	60～69	70～79	80～89	≥90
人数	1	5	2	8	14	37	14	8

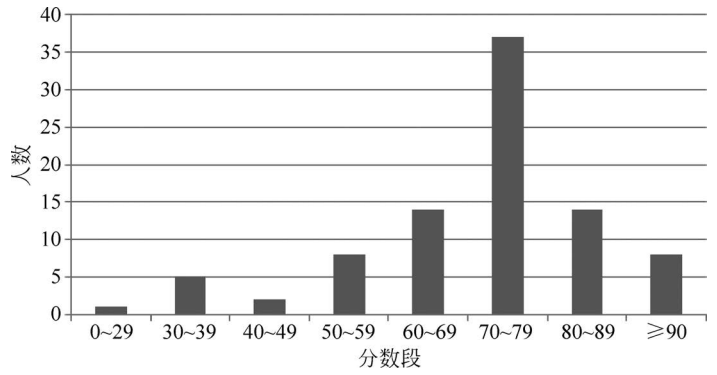


图 3-19 直方图

除了常用的图表外,可供大家选择的还有如下 7 种。

(1) 漏斗图:适用于业务流程比较规范、周期长、环节多的流程分析,通过漏斗各环节业务数据的比较,能够直观地发现和说明问题所在。

(2)(矩形)树图:一种有效地实现层次结构可视化的图表结构,适用于表示类似文件目录结构的数据集。

(3) 热度图:以特殊高亮的形式显示访客热衷的页面区域和访客所在的地理区域的图示,用于显示人或物品的相对密度。

（4）关系图：基于三维空间中的点—线组合，再加以颜色、粗细等维度的修饰，适用于表征各结点之间的关系。

（5）词云：各种关键词的集合，往往以字体的大小或颜色代表对应词的频次。

（6）桑基图：一种由一定宽度的曲线集合表示的图表，适用于展现分类维度间的相关性，以分流的形式呈现共享同一类别的元素数量，如展示特定群体的人数分布等。

（7）日历图：顾名思义，以日历为基本维度的对单元格加以修饰的图表。

在制作可视化图表时，首先要从业务出发，优先挑选合理的、符合惯例的图表，尤其在用户层次比较多样的情况下，要兼顾各个年龄段或者不同认知能力的用户的需求；其次是根据数据的各种属性和统计图表的特点来选择，例如，饼图并不适合用作展示绝对数值，只适用于反映各部分的比例。对于不同图表类型，带着目的出发，遵循各种约束，才能找到合适的图表。