

第3章

常用可视化的多维分析

在数据分析中,通过各种可视化的图形,从多个维度、多个层次展示企业商务的执行情况,发现可能存在的问题或潜在的危机,并预测未来业务发展的趋势,具有重要的价值。而且利用可视化的工具,也可以发现数据的一些质量问题、分布特点,可以为进一步的数据挖掘做预处理。因此,结合业务理解,利用常用的可视化工具,如 Tableau、Lumira、国内的永洪大数据分析工具 Yonghong Z-Suite 等,对数据做一定深度的分析,这是数据分析师的基本功。

可视化图形通过位置、长、宽、角度、大小、色调、形状等多个方面,以视觉效果来表达图形相关含义。在数据分析中,各种不同的图形具有各异的作用,这也就为解决多元问题,深层次了解业务逻辑提供了方法途径。

可视化图形的作用各异,箱图的作用是展现数据的离散状态,以其数据节点:上限、下限、上四分位、下四分位、中位数及异常值为依据,来分析数据的离散程度等信息,可应用于数据预处理,识别数据异常值及分析数据离散状态。雷达图的作用是对事物的不同维度进行分析研究(通常,维度应大于或等于四维),通过网状结构的图形对比形象展示各维度属性的相关状态。标签云的作用是显示词频,将标签出现或者被引用的多少,通过标签字体的大小和颜色等视觉效果呈现出来。气泡图的作用是研究数据之间的关系,以气泡的位置和大小及颜色来表现变量之间的关联。树图的作用是展现数据的层次关系,通过树图区域模块的占比、颜色深浅及层次等信息来研究数据之间的逻辑结构关系。地图的作用是展示数据与地理位置之间的关系,同时,可以根据颜色的深浅来判断地理区域或关键词的热门程度。高低图的作用是展现数据的波动特性,其不仅能研究数据长期波动的特性,也能研究数据短期的波动特性。双轴图的作用是展现数据的波动特征以及其数据之间的关联,通过在同一分析图形中绘制不同类别的图形,形象地展示数据之间的关系。关系图的作用是展现事物之间的相关性及其逻辑结构,以事物之间连线的粗细和颜色深浅等视觉效果为依据,研究事物之间复杂的逻辑关系。热图的作用是表现数据的热点特征,以视觉化的区域和色彩来表现数据的热点程度等特征。

在种类繁多的分析图形中,需要根据研究问题的不同,选取适当的分析图形来进行数据分析。下面对常用的分析图形作简要介绍。

3.1 箱图

箱图是一种显示数据离散状态的分析图形。通过箱图,能够获得相关数据节点等信息。箱图主要包含 6 个数据节点: 上限、下限、上四分位、中位数、下四分位和异常值。

通常来说,上限位于上四分位加上 1.5 倍四分位距处,下限位于下四分位减去 1.5 倍四分位距处(如果样本数据在上四分位加上 1.5 倍四分位距及下四分位减去 1.5 倍四分位距处无数据,则上下限即为样本数据的最大值和最小值,这种情况下就没有异常值了)。在箱图中,异常值定义为数据点在样本数据中的位置大于上四分位加上 1.5 倍四分位距或小于下四分位减去 1.5 倍四分位距的数据。其中,温和异常值使用“o”表示,极端异常值使用“*”表示。

在箱图中,箱子占据了样本数据的一半,因而,箱子的宽度在一定程度上反映了数据的波动程度。箱子中间的一条线代表了中位数,其反映了样本数据的平均水平,同时,当中位数偏离上四分位和下四分位中心位置时,数据就表现出一种偏态性,中位数越偏离箱子中心位置,偏态性越强。箱图的另一主要功能是识别数据异常。进行数据分析时,异常数据可能会对分析结果造成影响,因而,通过箱图识别出异常值,并将其剔除,这将有损于数据分析结果的正确性。与其他统计图形相比,箱图可以将多批数据放在同一坐标轴上,并排排列进行对比,使得样本数据特征的分析变得更加容易。

为了更加形象地了解箱图相关结构特点及功能,结合香水实例,使用 SPSS Statistics 工具绘制出的箱图如图 3.1 所示,统计分析香水价格的相关情况。

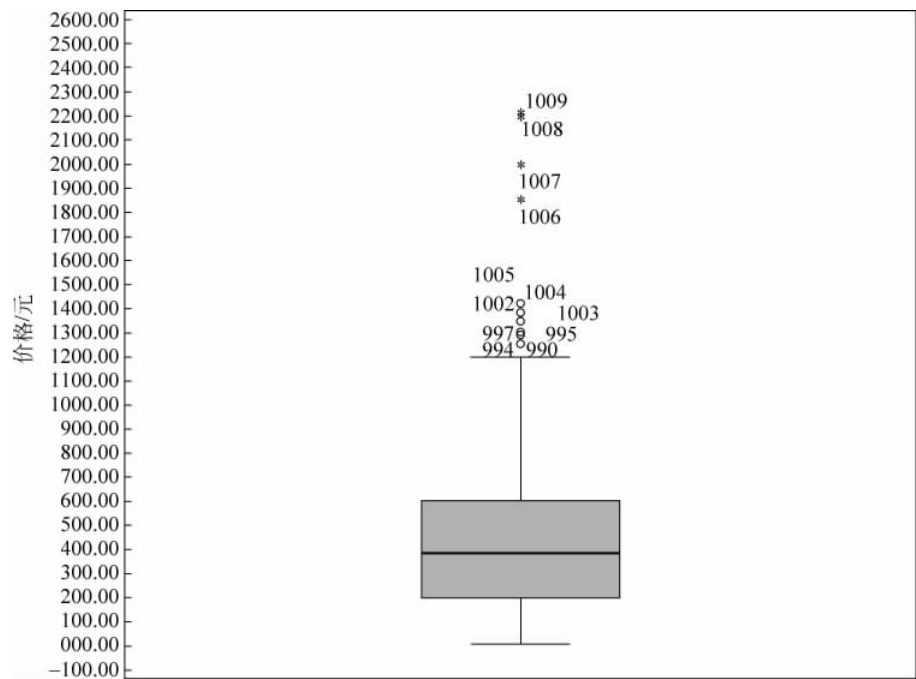


图 3.1 香水价格的箱图

图 3.1 所示箱图有关的几个数据如下。

下限：9.9，由箱子下方的一条线表示，下限由下四分位数减去 1.5 倍四分位距决定。下四分位：200，由箱子的下边线表示，代表数据的下四分位。中位数：385，由箱子中间的一条线表示，代表数据的中位数，反映了香水价格的平均水平为 385。上四分位：600，由箱子的上边线表示，代表数据的上四分位。上限：1189，由箱子上方的一条线表示，上限由上四分位数加上 1.5 倍的四分位距决定。

从图 3.1 中可以看到大于上限的圆圈点，这些点就是异常值，分析数据时可将其忽略。此外，这些数据点对应的标号是这些异常点在样本数据之中的位置，可以根据这个位置信息找到该异常点在原始数据中的具体位置。994~1005 号数据都是温和异常值，用“o”来表示；1006~1009 号数据都是极端异常值，用“*”表示。

绘制箱图前，有可能需要对数据进行预处理。举例说明：针对香水样本数据，评价量在一定程度上反映销售量，对“探究不同品牌香水评价量相关特征”这一问题进行分析，在未对数据进行预处理之前，使用 SPSS Statistics 工具绘制箱图，如图 3.2 所示。

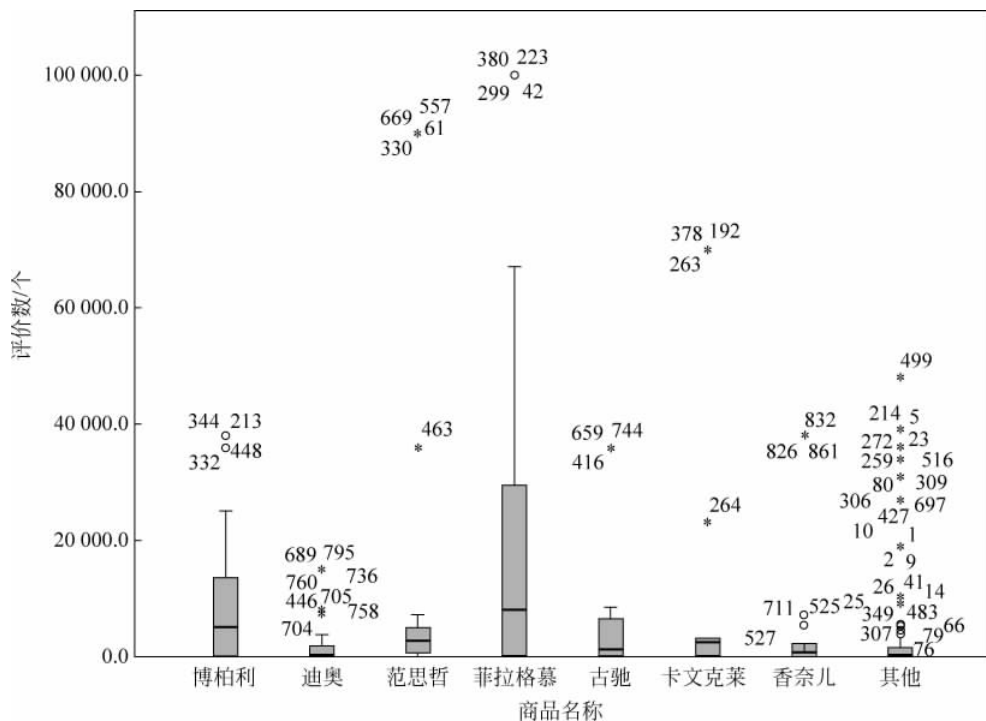


图 3.2 不同品牌香水的评价量箱图

由于箱图本身形状怪异，箱子被压扁且有很多的异常值，因而很难从图 3.2 中得到具体结论。分析其原因，是因为没有对样本数据进行预处理（当然，也不是所有样本数据都需要进行预处理）。针对此类问题，如果数据取值为正数，一个解决方法就是尝试使用对数变换来对数据进行预处理，使幂函数或指数函数的曲线拟合线性化，能够很好地处理不对称分布、非正态分布和异方差等情况。

针对本实例，首先使用对数变换来对样本数据进行预处理。使用底为 10 的对数进行处理，得到评价的对数变换结果，存储至评价数量这一变量中，然后绘制对数变换后不同品牌

香水的评价量箱图,如图 3.3 所示。

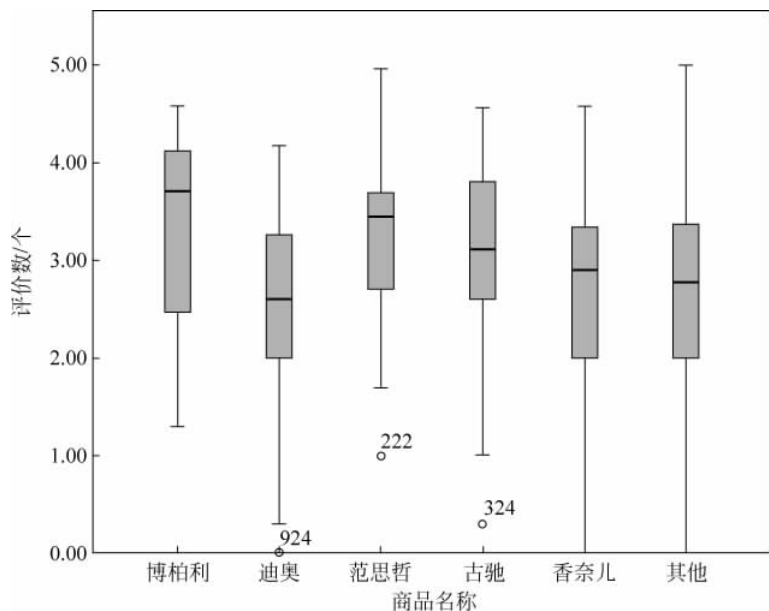


图 3.3 对数变换后不同品牌香水的评价量箱图

对比图 3.2 与图 3.3,发现经过对数变换的箱图可更加直观地表现出数据的平均水平、波动程度和偏态等信息。这一点说明不是所有数据都适合画箱图,但是可以利用数据变换进行预处理,使得数据适合用来绘制箱图。

箱图的另一功能是使用定性变量画分组箱图,各个箱图之间作比较。结合前面分析的实例,图 3.1 只设置了一个定量变量,所以只有一个箱图,而这就让箱图失去了它的一个很重要的功能:多批次数据的对比。而且,一个箱子的箱图是没有必要的,完全可以由直方图来代替。图 3.3 所示箱图设置了定性变量——商品名称,通过商品名称这一定性变量,就能在一个箱图中绘制多个箱子,在同一水平上对各个箱子相关数据节点进行比较,得到多批次数据之间的关系。

3.2 雷达图

雷达图是一种应用于多维数据分析的图形,通过对多维数据进行分析,来探究问题的相关状态。雷达图主要应用于财务分析,其主要作用是将各项数据分析的数或比率,集中展现在一个圆形的图形或者正多边形上,以凸显各种数据比率情况。在财务分析中,雷达图主要用于分析企业经营状况——收益性、生产性、流动性、安全性和成长性的状况。

雷达图可以从静态和动态两个方面分析客户的财务状况。静态分析是将客户的各种财务比率与其他客户或者整个行业的财务比率作横向比较;动态分析是将客户现在的财务比率与以前的财务比率作纵向比较,就可以发现客户财务及经营情况的发展方向和变化。雷达图将纵向和横向的分析比较方法结合起来,综合计算客户的收益性、流动性、成长性、安全性及生产性等 5 类指标。

雷达图的样式是一种类似于蜘蛛网状的图形,主要通过将企业的各种经营比率值连接而成的不规则闭环折线图与各同心圆进行比较,得到企业经营态势的好坏。其中,同心圆最小圆代表同行业平均水平的一半或者最差水平,中心圆中等大小的圆代表同行业的平均水平或特定比较对象的水平,大圆代表同行业平均水平的 1.5 倍或最佳状态。然后,将同心圆等分为 5 个扇区,每个扇区指代一个维度,分别代表收益性、安全性、流动性、成长性和生产性指标区域。

上述介绍的财务分析只是雷达图的主要应用领域。在雷达图的通用性方面,其适用范围和规则简要许多,主要解决多维数据的分析,且每个维度都是可以度量的。一般来讲,雷达图的维度数目应大于等于 4。

结合香水案例数据,先对数据进行预处理,选取中国、美国、法国、意大利及英国 5 国,分析其销售量、品牌数、产品质量、平均评价数及平均价格 5 个维度相关特性,其中,各个维度数据代表其占样本数据相对应维度总体的比例。使用 SAP Lumira 绘制雷达图,如图 3.4 所示。

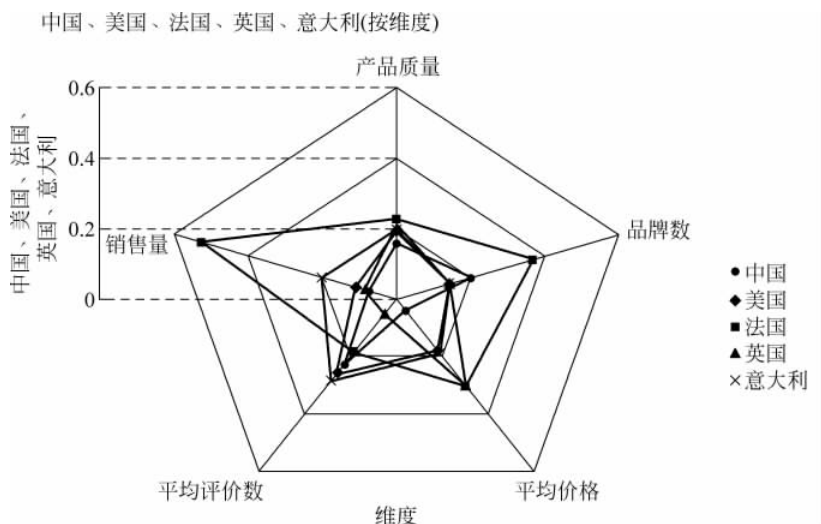


图 3.4 雷达图

分析图 3.4 所示的雷达图,这里的雷达图显然不是应用于企业的财务分析,而是雷达图通用性的体现。此图拥有 5 个维度,分别为销售量、品牌数、产品质量、平均评价数及平均价格。根据图 3.4 可以很直观地得到法国在 5 个维度上基本都占据了优势,因而可以推断出法国在香水领域处于领军地位,而这一点也符合我们对法国香水的认识。美国、英国及意大利 3 国在各个维度上水平相当,基本都处于中等水平。而中国香水除了在品牌数量之外,其余各个维度基本都处于较低层次水平,因而,中国品牌香水想在香水领域做出一番成就,还有很长的路要走。

综合以上分析,可以总结出以下几条:

- 雷达图主要应用于财务分析,对企业经营状况的 5 类指标(流动性、生产性、安全性、收益性和成长性)进行评价,来综合评估企业的经营状况。
- 使用雷达图之前,一般需要进行预处理,首先计算出所需分析维度的占比,然后进行

非常清楚地了解各品牌香水评价量之间的比较情况。虽然标签云能形象地表示数据之间的关系,但是其很难得到具体的数据情况。以直方图为例,分析各品牌销量之间的关系。香水品牌销量直方图如图 3.6 所示。

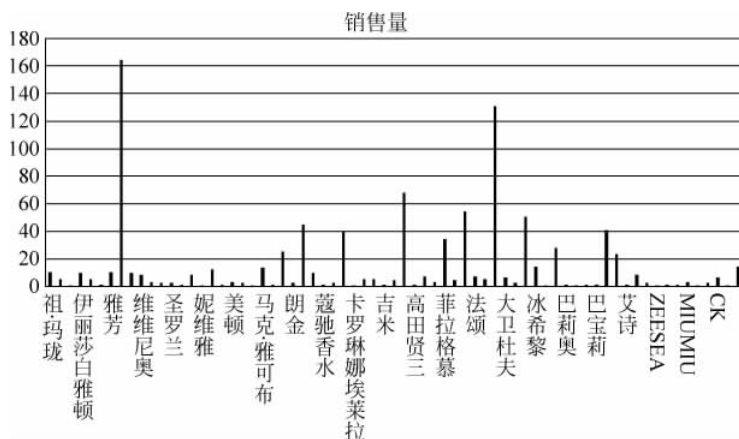


图 3.6 香水品牌销量直方图

对比图 3.5 和图 3.6,在不要求精确性的前提下,标签云比直方图更加形象地展现数据之间的关系,特别是针对权多的数据样本,绘制标签云更有利于对数据进行解读。

综合以上分析,总结如下:

- 标签云使用独立词汇,通过字号、颜色、排序和字体等属性,来形象地体现标签的使用次数及热门程度等相关特性。
- 标签云描述的特性并不能非常精确地呈现出来,因而标签云并不适用于那些对绘图结果要求非常准确的场景。

3.4 气泡图

气泡图中气泡的位置和大小由三维变量决定,其中第一组变量给出直角坐标系的 x 轴值,相邻组变量给出 y 轴值,第三组数据则指代气泡的大小,以上基本就是气泡图的逻辑构成。气泡图基本上与 XY 散点图类似,可以说,气泡图是散点图的升级。散点图只能对成组的两个数值进行比较,而气泡图可以对成组的多个数值进行比较。

使用香水案例数据,并对数据进行预处理,得到不同品牌的平均评价数量和销售数量以及平均销售价格。从中选取香奈儿、迪奥、兰蔻、古驰、范思哲、博柏利、安娜苏、爱马仕和卡尔克莱 9 个占据大多数销售量的品牌进行分析,对数据进行预处理,得到表 3.1。

表 3.1 品牌香水记录表

品牌	平均评价数	平均价格/元	销售量
香奈儿	4303	708	164
迪奥	1882	573	131
兰蔻	3073	388	45
古驰	6262	544	67

续表

品牌	平均评价数	平均价格/元	销售量
范思哲	9359	340	55
博柏利	9511	338	51
安娜苏	5262	289	41
爱马仕	848	717	24
卡尔克莱	12 251	319	40

通过表 3.1,以平均价格为 x 轴,以平均评价数为 y 轴,以销售量来确定气泡的大小,使用 SAP Lumira 绘制气泡图,如图 3.7 所示。

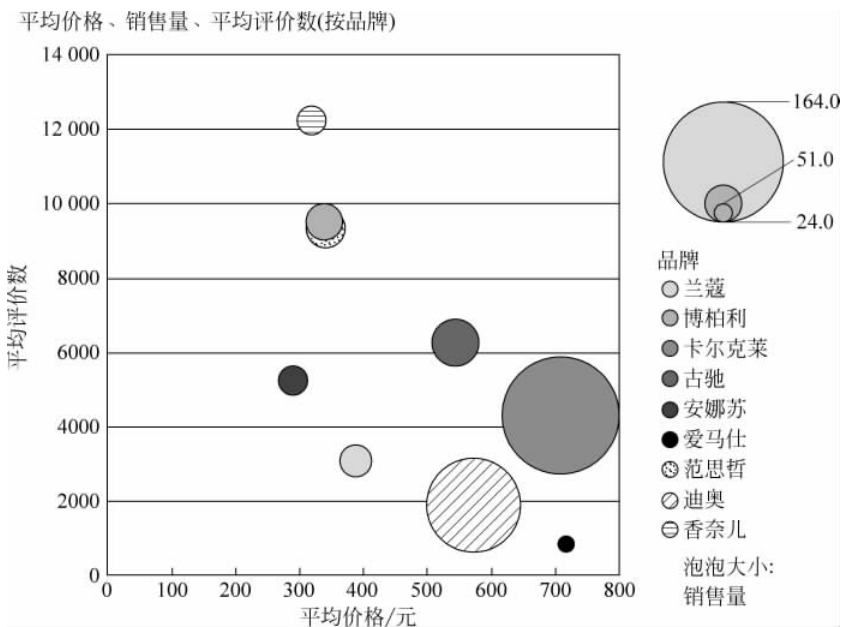


图 3.7 品牌香水气泡图

观察图 3.7 所示的气泡图,通过气泡的位置和大小,很直观地得到各品牌的平均评价数和平均价格及销量在总体样本数据中的具体情况。其中,在本例中,使用不同颜色标记气泡,每种颜色代表一种品牌,这样更利于对比分析各香水品牌的相关数据。气泡的大小反映了销量的多少。可见,香奈儿和迪奥品牌香水销量最高。

3.5 树图

树图是为了达到某种目的或者解决某一问题,采用目的方法或者结果原因方法,层层展开分析,以寻找最好的解决方法或者是查看其根本原因。树图从一个项目出发,展开两个或两个以上分支,然后从每一个分支再继续展开,以此类推,形似一棵树。

按照功能,可以将树图分为两类:对策型树图和原因型树图。对策型树图主要以目的方法方式展开,而原因型树图则以结果原因方式展开。

目的方法方式主要是遵循层层推进规则,对于每一目的,进行层层发问,寻求好的方法或者途径来达到目的,也就构建了对策型树图。结果原因方式则是针对结果进行发问,有哪些原因会导致这个结果或者哪些事项会对这个结果造成影响,通过这种层层推进分析,也就建立了原因型系统图。

树图还可分为矩形树图、组织结构图等。其中,组织结构图用于描述组织结构,一般采用自上而下的展开形式。而对于矩形树图,其主要用来展示层次关系数据。相对于其他层次图表,矩阵树图的优势在于更加直观,并且可以展示层级内的占比关系,直观地反映区域占总体的比率,同时,矩形树图还能依据区域颜色的深浅来反映不同关键词的热门程度。

树图通常用来将主要的类别逐渐分解成越来越详细的层,这样绘制树图有助于思维从一般到具体。在香水案例中,首先对数据进行预处理,得到不同品牌的销售量以及平均价格,并以销售量和平均价格为度量,品牌为维,使用 SAP Lumira 绘制矩形树图,如图 3.8 所示。

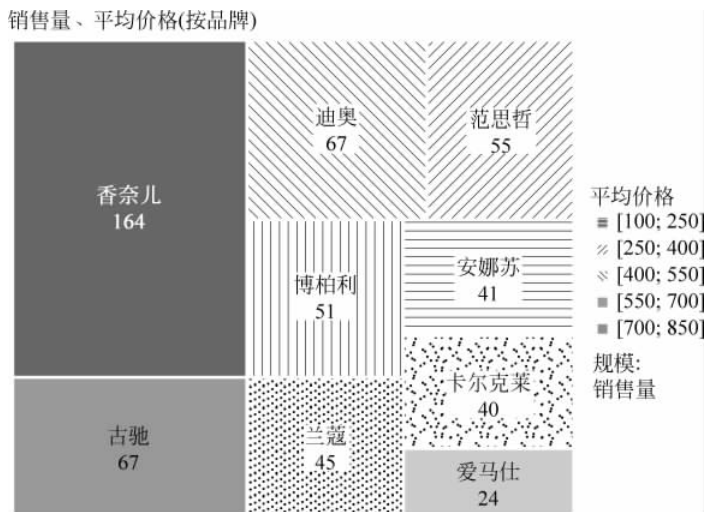


图 3.8 树图

图 3.8 所示树图就是一种典型的矩形树图,其中销售量为第一度量,根据销售量的多少来划分矩形大小。平均价格为第二度量,根据其平均价格的高低决定相应数据模块的颜色深浅。从图 3.8 可以非常直观地看出,香奈儿品牌香水不仅销量好,评价多,而且价格还比较高。这说明了香奈儿品牌香水口碑好,销售情况好。

3.6 地图

地图是以一定的数学、符号化、抽象化法则,使用制图方法,反映客观实际形象的符号模型或者图形数学模型。地图是按照一定的比例运用符号、颜色、文字注记等描述显示地球表面的自然地理、行政区域、社会经济状况的图形。在适用于数据分析的地图图形前提下,这

里介绍的是数据地图。遇到数据与地名的场景时,使用地图能更加形象地展示数据与地理位置之间的关系。数据地图是一种将数据与地理信息有机结合的一种地理数据表达方式。数据地图是以图形化的方式分析和展示与地理位置相关的数据,使得数据与地理之间的关系更加直观、形象化。地图能够很直观地反映数据与地理位置之间的关系,且这种关系可以是分层的,通过地图下钻操作,可以探查数据与不同层次地理位置之间的关系。同时,还能通过颜色的深浅来判断地区或关键词的热门程度。

地图图层是对空间表达的一种重要途径。一个地图可以拥有多个图层,将其叠加就能得到地图的底层(类似于背景图层),构成地图中最基本的地形、地貌数据及某些附属数据或信息。

建立地理智能对象时,如果将地理位置信息转换为地理层次结构,建立地理层次结构,就能实现地图的下钻操作。在香水案例中,以商品产地为例,对数据进行预处理后,建立地理层次结构。以评价量为度量(评价量能在一定程度上反映商品销量),使用 SAP Lumira 绘制包含下钻操作的地图。在地理维度里会出现分层次的地理维度选项,可以选择以国家或地区为维度标准,也可以选择以城市为地理维度标准,得到以城市为维度的地图。

地图下钻能够很好地处理地理位置之间具有层次关系的问题。在进行下钻操作后,可以将研究问题进行细化,分析问题的局部特点。例如,通过对中国香水销售的分析可以了解到中国香水的几个产地,以及该产地香水评价量的平均水平,进而推算出各产地销售量的大概水平。

3.7 高低图

高低图是采用多条垂直线段表示数值区域的统计图形,能够将数值区域形象地表示在图形上,通过多组数据并行比较,易于分析数据区域的相关特性。高低图与折线图、散点图、条形图等统计图相比,它既有研究数据长期变化的特性,也有研究短期内数据变化的特性。因为这些特性,高低图广泛应用于股票、商品、货币及其他市场数据分析中。

高低图绘制过程中,纵坐标表示一个三维数据,分别为高值、低值及关闭。高值代表对应数值区域的上限,低值代表对应数据区域的下限,关闭是用户指定的一个特殊变量(如数据集的平均值),可以在数据区域上以小圆圈的形式标出。

为了更加直观地了解高低图,依据香水样本数据,以香水价格平均数为关闭变量、香水价格最小值为低值变量、香水价格最大值为高值变量,以香水品牌确定类别,使用 SPSS Statistics 工具绘制高低图,如图 3.9 所示。

图 3.9 显示了一些主要的香水品牌价格的高低图,从中可以直观地了解各数据集的区域范围。通过关闭在高低图中小圆圈的表示,可以了解各品牌香水的平均价格。通过对高值变量的分析,可以得到迪奥和香奈儿两个品牌香水基本占据了高端香水市场。各品牌香水的平均价格在 400~700 元波动。

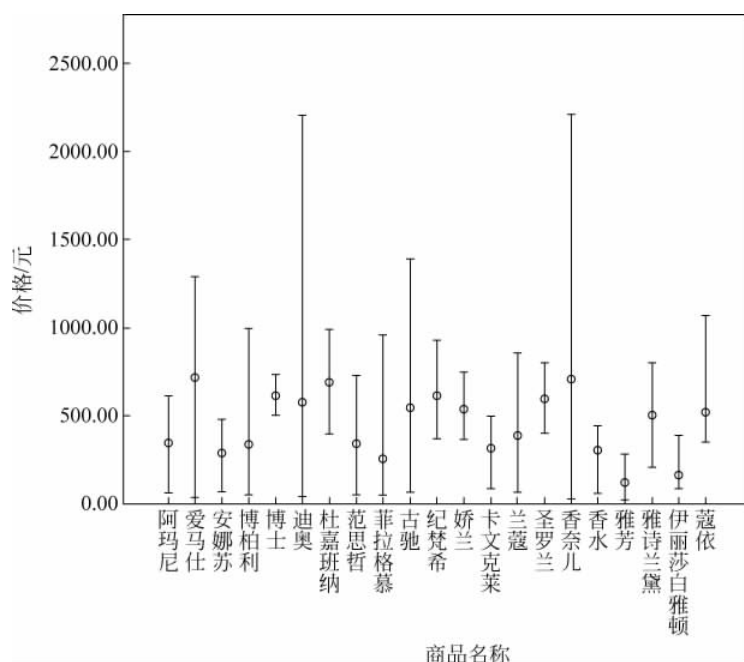


图 3.9 高低图

3.8 双轴图

双轴图是一种单 x 轴双 y 轴的统计图形, 这里的双 y 轴指有两条 y 轴, 位于两侧, 代表了两种变量含义, 这样就能根据 x 轴变量变化绘制出两条不同图形, 可以简易地看成两种统计图形的合并。绘制的图形可以采用多类图形, 如折线图、直方图和散点图等。

双轴图可以依据 x 轴的分类类型分为两类: 包含分类 x 轴的双 y 轴及包含刻度 x 轴的双 y 轴。分类 x 轴的双 y 轴类型对应的 x 轴变量是非连续型变量, 而包含刻度的 x 轴的双 y 轴类型对应的 x 轴变量是连续型变量。

双轴图能够在同一统计图形上采用两种绘图方式, 并且将结果展示在同一统计图形上, 使得能够更加形象地对比分析多组数据特征。当采用不同统计图形的绘制方法绘制双轴图时效果对比明显, 如折线与直方图的叠加, 就能够很好地展示数据特征。

为了更加形象地表示双轴图的特征, 依据香水案例, 设置香水品牌为分类 x 轴, 左侧 y 轴为各个品牌评价平均数, 右侧 y 轴为各个品牌价格平均数, 使用 SPSS Statistics 工具绘制的双轴图如图 3.10 所示。

通过图 3.10 所示的双轴图, 在已列的各品牌香水中, 香水的平均评价数高的对应的平均价格一般较低, 香水的平均评价数低的对应的平均价格一般较高。由此可见, 不同品牌香水的平均评价数与平均价格呈现一种负相关趋势。对于那些平均评价数中等的品牌, 其平均价格也保持中等。

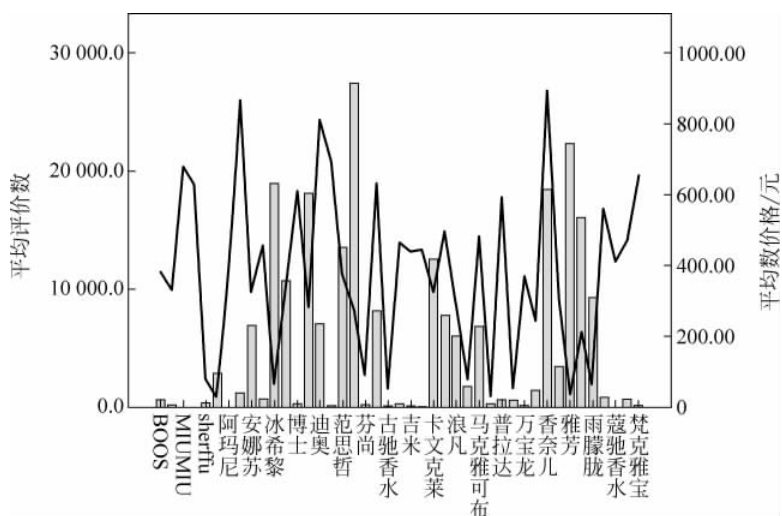


图 3.10 双轴图

3.9 关系图

关系图是使用连线的方式,将相关事务连接起来,表示事物相关性的图形,这里的相关性指的是事物之间复杂的逻辑关系。关系图通过对事物相关性的研究,找出事物之间复杂的逻辑关系,从而找出要素和解决问题的方法。

在对事物的逻辑结构进行分析时,如果分析的关系是纵向关系,即分析的是单个事物内部因素之间的关系,则可以选用“原因结果、目的方法等”方法来研究。如果分析的关系是横向关系,即需要分析多个事物之间的复杂逻辑结构,这时就需要应用关系图。

关系图按其应用目的来分,可分为:单一目的型和多目的型(研究问题的个数);按照其分布结构来分,可分为:中央集中型(箭头向内集中)和单向汇集型(箭头单向顺延)。在对关系图分析时,箭头只进不出的是问题;箭头只出不进的是重要因素;箭头有出有进的是中间因素。需要指出的是,一般适用于关系图的场景应该是事物之间的逻辑关系非常复杂的,简单的场景可以由树图等来研究其相关逻辑特征。

为了更加直观地了解关系图的相关特征,结合香水案例,对商品名称、香调、分类、商品产地及性别 5 个字段使用关系图来了解其相互逻辑关系。使用 IBM SPSS Modeler 绘制关系图,如图 3.11 所示。

分析图 3.11 所示的关系图,关系图中各元素字段之间使用线段粗细来表示两因素之间联系的强弱,线段越粗、颜色越深,代表其联系越强。这里给出了其链接强度的相关数据,见表 3.2。

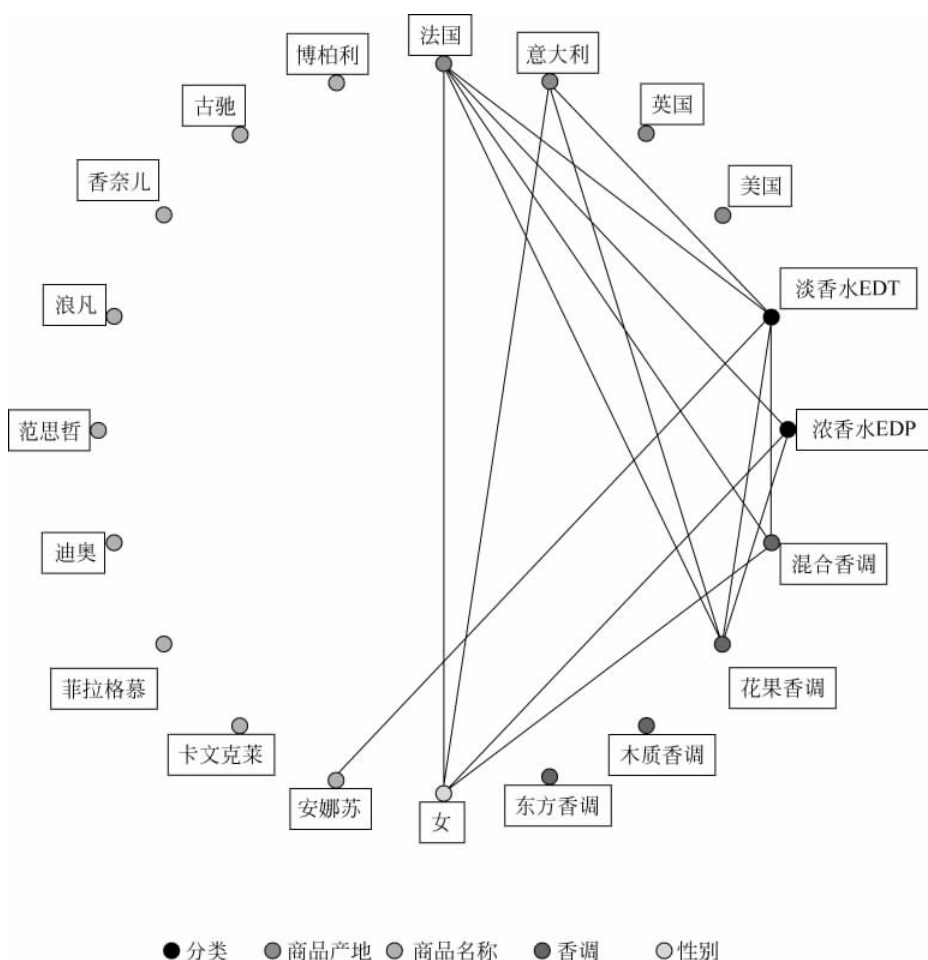


图 3.11 关系图

表 3.2 链接强度表

链接	字段 1	字段 2
6.65%	香调="花果香调"	性别="女"
5.98%	分类="淡香水 EDT"	性别="女"
4.66%	商品产地="法国"	性别="女"
4.62%	香调="花果香调"	分类="淡香水 EDT"
3.27%	分类="浓香水 EDP"	性别="女"
3.21%	香调="花果香调"	商品产地="法国"
2.76%	分类="淡香水 EDT"	商品产地="法国"
2.12%	香调="花果香调"	分类="浓香水 EDP"
1.99%	分类="浓香水 EDP"	商品产地="法国"
1.92%	商品产地="意大利"	性别="女"
1.67%	香调="混合香调"	性别="女"
1.58%	分类="淡香水 EDT"	商品产地="意大利"
1.54%	香调="花果香调"	商品产地="意大利"

续表

链接	字段 1	字段 2
1.03%	商品名称="古驰"	性别="女"
0.98%	香调="混合香调"	分类="淡香水 EDT"
0.98%	香调="混合香调"	商品产地="法国"
0.94%	商品名称="博柏利"	商品产地="法国"
0.92%	商品名称="博柏利"	性别="女"
0.85%	商品产地="美国"	性别="女"
0.83%	商品名称="古驰"	商品产地="法国"
0.79%	商品名称="范思哲"	性别="女"
0.77%	商品名称="博柏利"	香调="花果香调"
0.75%	香调="花果香调"	商品产地="美国"
0.75%	商品名称="范思哲"	商品产地="意大利"
0.75%	商品名称="古驰"	香调="花果香调"

对照表 3.2 和图 3.11,可以看出香水香调为花果香调、香水分类为淡香水 EDT、性别为女的链接关系占据高位,法国的香水也与女性的链接关系比较高。这表明了女性用户更加青睐法国产花果香调淡香水 EDT 类型的香水,而且花果香调的香水与淡香水 EDT 类型更加搭配。

3.10 热图

热图是一种表现数据热点的图形,以区域和颜色等视觉效果,形象地表现数据的密度、频率及热点等特征。热图的热字体现在图形表达时,其数据热度等信息一般以火焰色彩表示,展现出极强的视觉表达力。

热图是以区域颜色深浅效果来展现数据特征,因而,热图表达的仅仅是数据之间的大概关系,并不能精确展现数据频率、热度等特征。

热图可以看作是地图的增强版。地图展现的是数据与地理位置之间的关系,而热图则是将地理位置广义化,以不同的区域块来区别。

在 Web 领域,热图被用来检测页面的哪些部分对顾客具有吸引力。其原理是:记录用户点击的区域,以热图的形式展现出用户得到点击区域,然后根据具体情况对页面设计进行调整,设计出更具用户友好型的网站。在其他领域,热图也有很广泛的应用,天气预报中的气温图就是典型的热图。

下面结合香水案例来了解热图。香水的评价量在一定程度上反映香水的销售量,以香水平均评价为度量,以商品名称及香调为维度,使用 SAP Lumira 来绘制热图,如图 3.12 所示。

分析图 3.12,该热图表现的是香水品牌的各香调类型香水的评价量(销售量)分布热点。各香水品牌的香调类型主要是花果香调类型和混合香调类型,其平均评价量热点分布主要是 0~15 000 的水平。其中,卡文克莱及香奈儿拥有大多数香水香调类型,且香奈儿香调类型香水的平均评价较高,而卡文克莱的平均评价较低,反映了香奈儿香水的销量高,卡文克莱的销售量较低。



图 3.12 热图