第 3 章

描述性多元分析

我们在联系人数据集中已经看到了，在现实生活中，属性的数量通常超过两个，可以是数十、数百个甚至更多。实际上，以生物学为例，具有数百甚至数千个属性的数据集是很常见的。当一个数据集的分析涉及两个以上的属性时，称为多元分析。与单元分析和双元分析一样，频数表、统计手段和图表可以用于多元分析。

因此，我们在第 2 章中为单元分析和双元分析描述的一些方法既可以直接使用，也可以修改为使用任意数量的属性。当然，属性的数量越大，分析就越困难。必须要注意的是，用于两个以上属性的所有方法也可以用于两个或一个属性。

为了说明本章中描述的用于多元分析的方法，如表 3.1 所示，我们向第 2 章私人通讯录数据集添加一个新的属性。由于这个表有 7 列，我们的多元分析最多可以使用 7 个属性。列（属性）包括联系人、他们的家乡前一个月的最高温度记录、体重、身高、认识他们的时间（年数）以及性别，最后是对关系的评价。

接下来，我们将介绍第 2 章 3 种数据分析方法（频数、可视化和统计）中的简单多元方法，以及展示如何将它们应用于这个数据集。

表 3.1　包含身高和体重的私人通讯录数据集

联系人	最高温度/℃	体重/kg	身高/cm	年数	性别	关系
Andrew	25	77	175	10	男	好
Bernhard	31	110	195	12	男	好
Carolina	15	70	172	2	女	差
Dennis	20	85	180	16	男	好
Eve	10	65	168	0	女	差
Fred	12	75	173	6	男	好
Gwyneth	16	75	180	3	女	差
Hayden	26	63	165	2	女	差
Irene	15	55	158	5	女	差
James	21	66	163	14	男	好
Kevin	30	95	190	1	男	差
Lea	13	72	172	11	女	好
Marcus	8	83	185	3	女	差
Nigel	12	115	192	15	男	好

3.1　多元频数

每个属性的多元频数值都能独立计算。可以用一个矩阵表示每个属性的频数值，其中的行数是属性假设的值的数量，列数是频数值，表 2.3 中属性“身高”就是这样的一个例子。

第 2 章已经介绍过了，根据属性值是离散的还是连续的，分别用概率质量函数或概率密度函数定义属性值。定性和定量尺度使用不同的方法，不过，对于每个属性，可以采取以下频数度量。

- 绝对频数
- 相对频数
- 绝对累积频数
- 相对累积频数

3.2　多元数据可视化

我们已经看到，对于单元分析和双元分析，使用可视化技术表示数据和实验结果理解起来更容易一些。但是，第 2 章中涉及的大部分图表不适用于两个以上的属性。

好消息是，前面的一些图表可以进行扩展，以表示少量的其他属性。此外，新的可视化方法和技术不断被创造出来，以处理新的数据类型、解释结果的新方法和新的数据分析任务。根据属性的数量以及表示数据的空间和/或时间的需要，可以使用不同的图。本节探讨如何以不同的方式可视化多元数据，以及这些替代方法的主要优点。

若多元数据有 3 个属性，或者只能从一个多元数据集分析 3 个属性时，仍然可以在二元图中显示数据，将第 3 个属性的值与图中每个数据对象的表示方式联系起来。如果第 3 个属性可量化，则该值可以用图中对象的大小表示。

例 3.1　如图 3.1 所示，图中每个对象的大小与该对象的第 3 个属性数值成比例。

另外，如果第 3 个属性是定性的，则它的值可以在图中表示为物体的颜色或形状。而颜色或形状的数量将是属性可以假定的数值数量，在分类任务中，一般利用颜色和形状表示类标签。

图 3.2 中有两个图，其中第 3 个属性是定性的。右图将每个定性值表示为不同的形状，左图则用不同的颜色表示每个定性值。

另一种表示 3 个属性的方法是使用三维图，其中每个轴关联一个属性。如果这 3 个属性是定量的，那么这种方法就更有意义，因为相应属性的值可以在每个轴上表示，假设它们满足某种排列顺序。图 3.3 展示了一个使用联系人数据集的 3 个定量属性的三维图示例。

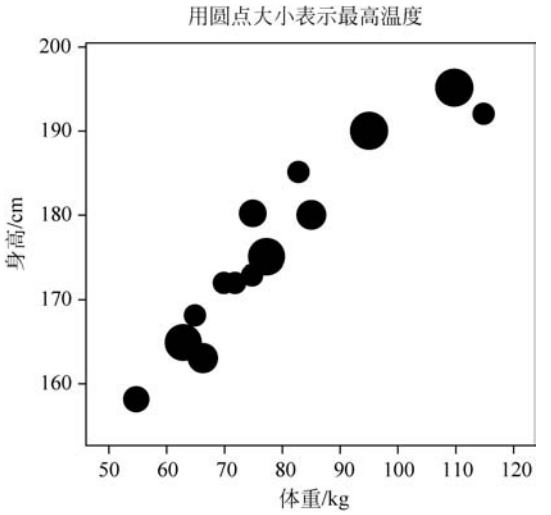

图 3.1　具有 3 个属性的对象图

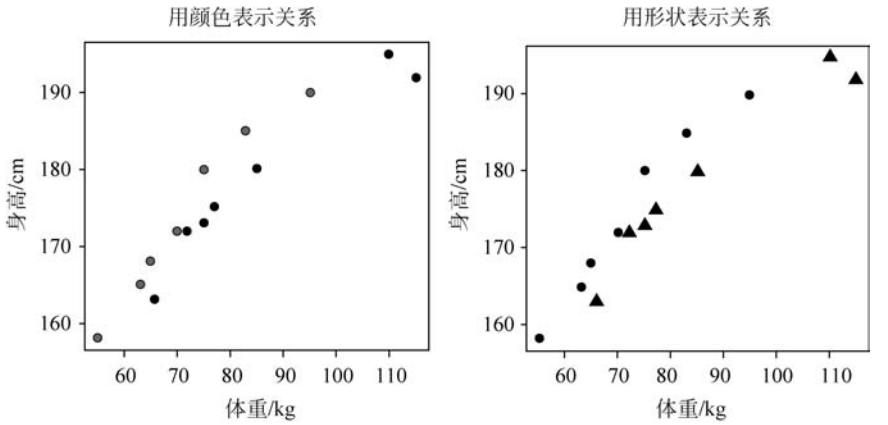

图 3.2　三属性图的两种形式，其中第 3 个属性是定性的

有人可能会问，3 个以上属性之间的关系应该如何表示？一种直接的方法是修改图 3.3 中的三维图形，通过所画对象的大小、颜色或形状表示第 4 个属性。图 3.4 中的属性使用不同的颜色表示。

尽管我们也可以使用三维图以及不同的格式和颜色表示前面图中两个以上的预测属性，但是并不是所有的图都允许这样做，或者，当这么处理时，得到的图可能会非常混乱。例如，根据所选的图，颜色和形状无法保留定量值的原始顺序和大小，只有不同的值。此外，一些定性的值也不会自然地由不同的对象大小来表示。

因此，若属性超过 4 个，就应该使用不同的图。还有一些图专门针对两个以上的定量属性，它们通常描述数据集的数量属性。其中最流行的是平行坐标图，也称为剖面图。数据集

中的每个对象都由一组穿过若干等距平行垂直轴的行序列表示，每个轴代表一个属性。将每个对象的线连接起来，然后用连续的直线表示每个对象，这些直线具有向上和向下的斜率。对于特定对象，这些线和垂直轴相连，且其位置和轴相关的属性值成比例，属性值越大，则位置越高。

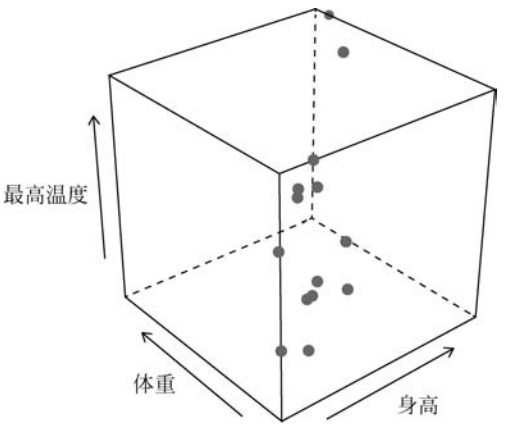

图 3.3　联系人数据集中的三属性图

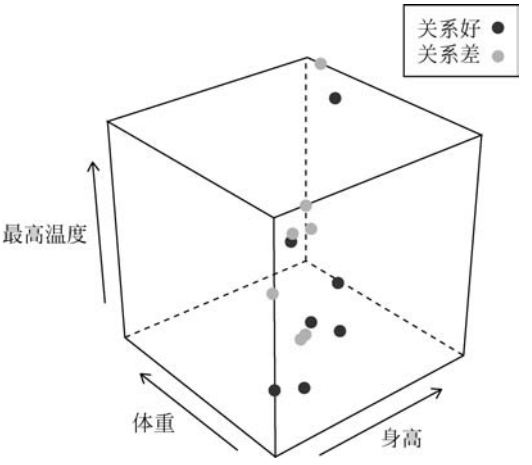

图 3.4　联系人数据集中的四属性图，第 4 个属性用不同颜色表示

例 3.2　图 3.5 是 4 个联系人的平行坐标图，使用了 3 个定量预测属性。数量属性在垂直轴上与其值相关的位置。每个对象都由一组线表示，这些线以表示属性值的高度穿过垂直轴，很容易就能看到每个对象的属性值。从图 3.5 中可以看出，其中 3 个对象的属性值具有相似的模式，这与第 4 个对象的属性值有很大的不同。图 3.5 中还显示了每个属性的最小值和最大值，也就是每个垂直轴上的最大值和最小值。

如图 3.6 所示，当我们添加更多的对象和属性时，这些线会相互交叉，分析的难度也会加大。还可以看出，定性属性可以用平行坐标图表示。此时纳入了定性属性“性别”，由于它只有两个值：男和女，所以所有对象都位于垂直轴上的两个位置之一。

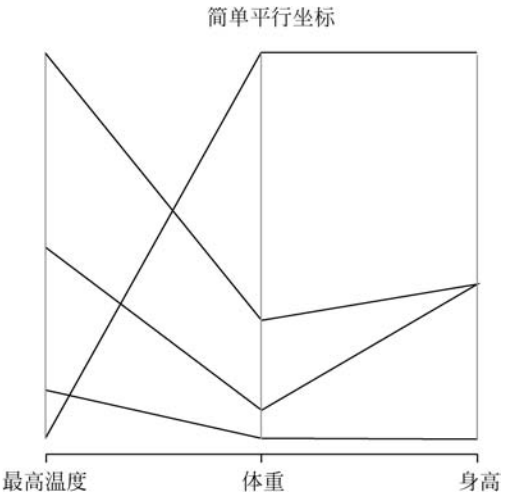

图 3.5　三属性平行坐标图

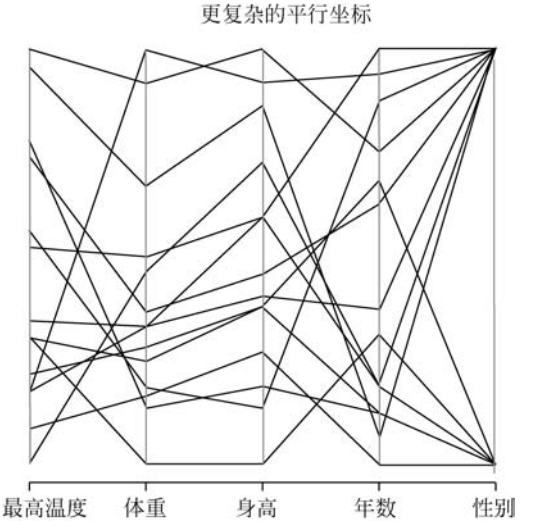

图 3.6　五属性平行坐标图

尽管平行坐标图看起来很混乱，但是我们可以为每个类分配一种颜色或样式，绘制相应对象的线条，从而简化对平行坐标图中的数据分析。因此，相同类对象的线条序列的颜色或样式也是一样的。图 3.7(a)对前一个图进行了修改，使用实线表示关系好的联系人，虚线则表示关系不好的。即使做了这样的修改，对图中的信息进行分析也是相当困难的。

这些图解释起来是否容易，取决于所使用的属性的顺序。如果来自不同对象的线条不断交叉，就很难从图中提取信息，改变绘制属性的顺序则可以减少交叉。如图 3.7(b)所示，水平轴上属性的顺序变化了，理解起来也稍微容易些。

平行坐标图中的每条线代表一个对象，若对象不多，且希望对它们逐个查看，那么可以

使用另一个图，也就是星图(也称为蛛网图或雷达图)。图 3.8 所示为包含了 4 个定量属性(最高温度、身高、体重和年数)的星图。为了避免图中属性值较大的情况，所有属性的值都统一为[0.0，1.0]区间。当一个属性的值接近 0.0 时，其对应的星将由于太靠近中心而无法被看到。这一点在标有“Irene”的星上很明显，从表 3.1 可以看出，Irene 的一些属性值最小。

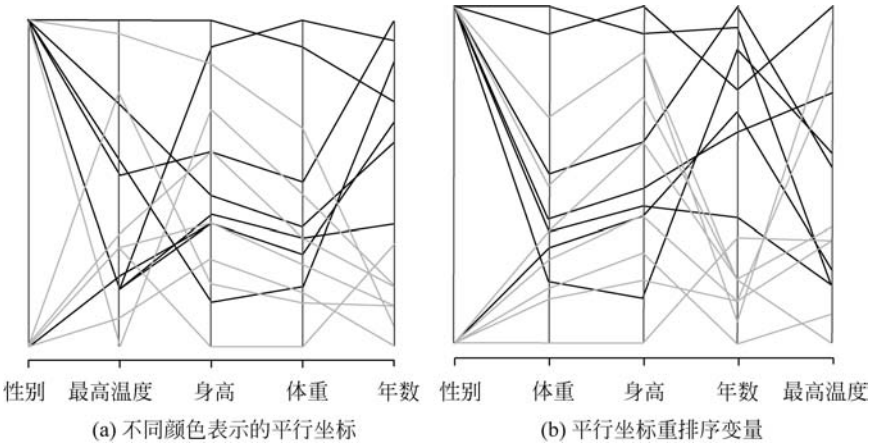

图 3.7　多属性平行坐标图

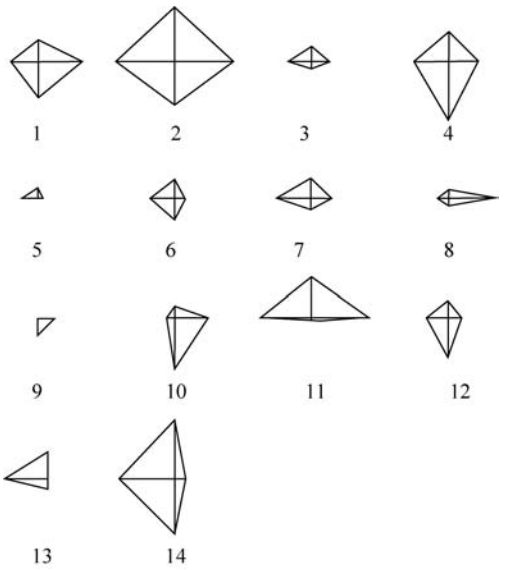

图 3.8　联系人数据集中每个对象的各属性数值的星图

定性属性也可以用星图表示，但由于它们的值很少，定性属性的点的变化很少，我们也可以在星图中标记每颗星。图 3.9 显示了 5 个属性(最高温度、身高、体重、年数和性别)的两个星图，混合了定量和定性属性。在图 3.9(a)中，每颗星都用联系人的名字进行了标记；在图 3.9(b)中，每个对象则都用自己的类标记。

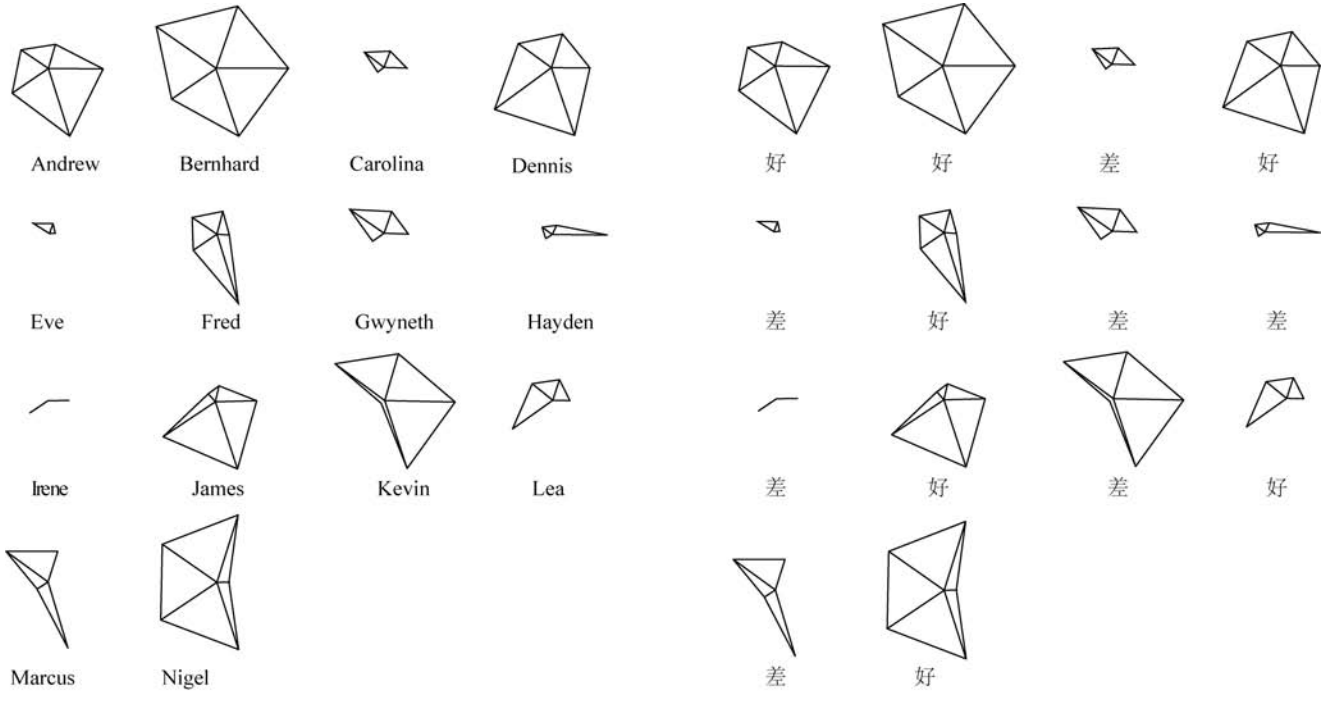

图 3.9　联系人数据集中每个对象的各属性数值的星图

即使有了这些改动，使用星图识别联系人属性值之间的差异的方式仍然不是很直观。利用人类识别人脸的能力，Herman Chernoff 提出了使用人脸表示对象，这种方法现在称为 Chernoff 脸谱。每个属性都与人脸的不同特征相关联，如果属性的数量小于特性的数量，则每个属性可以与不同的特性相关联。图 3.10 显示了如何使用 Chernoff 脸谱表示联系人数据集，其中使用了属性“最高温度”“身高”“体重”“年数”和“性别”。

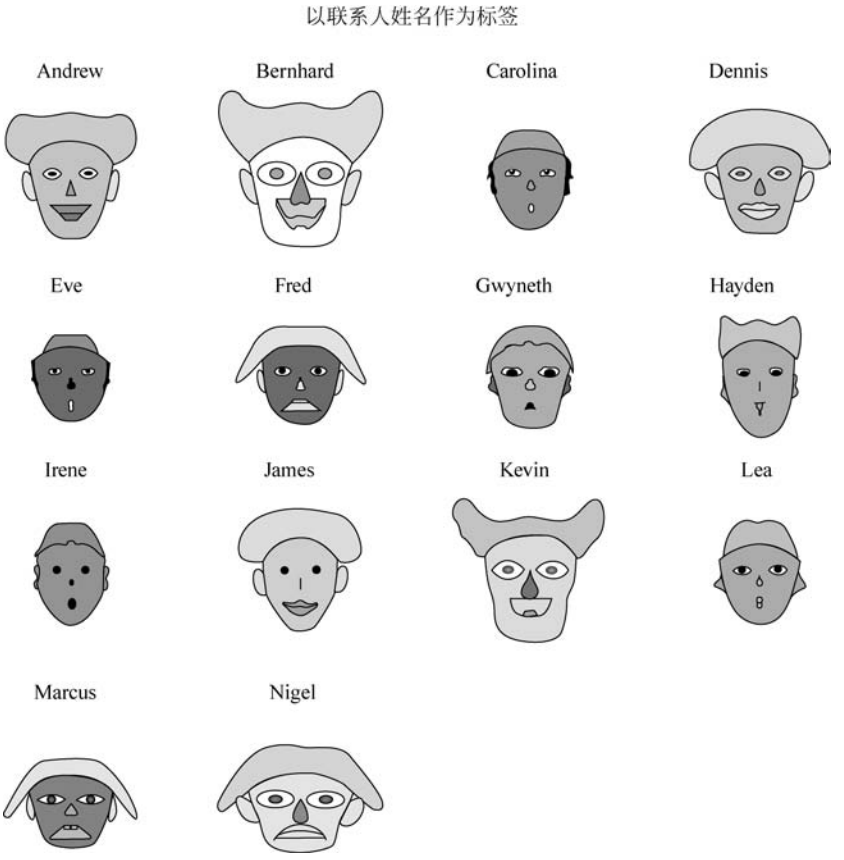

图 3.10　利用 Chernoff 脸谱将联系人数据集形象化

Chernoff 脸谱对于聚类也很有用，我们将在第 5 章中看到，它们可以用来说明每个聚类的关键属性。

还有其他几个图也很有用，可以从不同角度查看数据集。例如，可以使用流图查看数据分布如何随时间变化。数据可视化是一个非常活跃的研究领域，并在最近几十年得到了迅速发展，使数据分析更简单和更全面的新方案也在不断出现。

交互式可视化绘图的开发和使用是这一领域的一个趋势，用户与绘图进行交互使可视化信息更有用。例如，用户可以利用三维图形查看位置。有关数据可视化的进一步信息，建议读者查阅数据可视化方面的文章。

3.3 多元统计

乍一看，从两个以上的属性中提取统计度量似乎很复杂。不过，多元统计只是第 2 章所述的单元统计的一个简单扩展。我们将看到，以前描述的单元分析和双元分析的一些统计度量，如均值和标准差，可以很容易地扩展到多元分析。

3.3.1 位置多元统计

要测量有多个属性时的位置统计信息，只须测量每个属性的位置。每个属性的多元位置统计值都可以独立计算，这些值可以用一个元素数量等于属性数的数字向量表示。

例 3.3 举一个超过两个属性的位置统计的例子，表 3.1 中 4 个属性为“最高温度”“身高”“体重”和“年数”，它们的位置数据在表 3.2 中表示为一个矩阵，其中每行都有统计测量的 4 个属性。要使用标准格式，所有值都用实数表示。

表 3.2 定量属性的位置多元统计

位置多元统计	最高温度/℃	体重/kg	身高/cm	年数
最小值	8.00	77	175	0.00
最大值	31.00	110	195	16.00
均值	18.14	70	172	7.14
众数	15.00	85	180	2.00
第一四分位数	15.25	65	168	2.25
中位数或第二四分位数	15.50	75	173	5.50
第三四分位数	24.00	84.5	183.75	11.75

在第 2 章我们看到了一个单元分析的简单图形（箱形图）也可以用来呈现相关信息的多元数据集的属性。如果属性的数量不是太大，每个属性都可以使用一组箱形图。

例 3.4 图 3.11 列出了我们从联系人数据集中摘录的定量属性的等效结果，可以看到这些属性的值是如何变化的。从箱形图可以看出，“体重”属性的数值间隔比“年龄”属性的间隔大，并且“体重”属性的中位数比“最高温度”属性的中位数更接近数值的中心。必须要注意的是，为定性数据集绘制箱形图没有意义，因为除了众数之外，其他统计数据只适用于数字型数据。

若属性数量过大，如 10 个以上，则很难分析所有箱形图包含的信息。

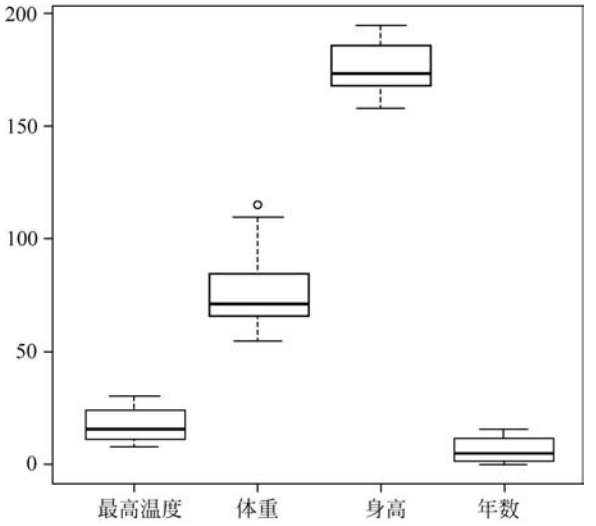

图 3.11　超过一个变量的箱形图

3.3.2　离散多元统计

对于多元统计，如第 2 章中介绍的振幅、四分位数范围、平均绝对偏差和标准差等离散统计量，可以对每个属性进行单独定义。

例 3.5　表 3.3 列出了表 3.1 数据集的属性“最高温度”“身高”“体重”和“年数”的多元离散统计示例。与多元位置统计的例子一样，离散统计可以显示在一个矩阵中，4 行中的每行代表 4 个属性的统计度量。

表 3.3　定量属性的离散多元统计

离散多元统计	最高温度	体重	身高	年数
振幅	23.00	60.00	37.00	16.00
四分位数范围	11.75	17.50	14.75	9.50
$\overline{\text{MAD}}$	7.41	14.09	11.12	6.67
s	7.45	17.38	11.25	5.66

前面描述的统计对每个属性的离散度进行独立度量，我们还可以度量一个属性的值与另一个属性的值之间有何不同。例如，如果属性 A 的值逐渐增加，属性 B 也会增加吗？如果是这样，我们说它们有相似的变化，所以一个与另一个成正比。如果变化方向相反（当属性 A 增加时，属性 B 减少），我们说这两个属性变化相反，它们是成反比的。如果没有观察到这两种情况，那么两种属性间可能就没有关系。

如 2.3.1 节所述，两个属性之间的关系使用协方差或相关性进行评估。一组属性中所有属性对的协方差度量都可以用协方差矩阵表示，在这些矩阵中，属性在行和列中的顺序相同。

例 3.6 在表 3.4 中，我们看到了联系人数据集中 4 个属性的协方差矩阵。每个元素表示一对属性的协方差，这很好地解释了数据集的离散情况。矩阵的主对角线表示每个属性的方差，这个矩阵也是对称的，因为主对角线上面与下面的值相同。这说明在计算协方差时，属性的顺序是不相关的，也可以看出体重和身高有很大的协方差。

表 3.4 定量属性的协方差矩阵

属性	最高温度	体重	身高	年数
最高温度	55.52	34.46	20.19	5.82
体重	34.46	302.15	184.62	42.39
身高	20.19	184.62	126.53	14.03
年数	5.82	42.39	14.03	31.98

例 3.7 在表 3.5 中，我们可以看到每对属性是如何关联的。我们使用联系人数据集中的 4 个定量属性。在 Pearson 相关矩阵中，每个元素显示一对属性的 Pearson 相关性。矩阵主对角线上的值都等于 1，表示每个属性都与自身完全相关。

表 3.5 定量属性的 Pearson 相关矩阵

属性	最高温度	体重	身高	年数
最高温度	1.00	0.27	0.24	0.14
体重	0.27	1.00	0.94	0.43
身高	0.24	0.94	1.00	0.22
年数	0.14	0.43	0.22	1.00

在第 2 章中，我们了解了如何绘制两个属性之间的线性相关性。我们可以使用类似的图说明一组属性中所有属性对间的相关性，这些属性利用几个散点图的矩阵，每对属性使用一个散点图。与相关性类似，散点图可以应用于任意数量的有序或定量属性对，以创建散点图矩阵，也称为窗格图。

例 3.8 图 3.12 所示为表 2.1 联系人数据集中所有属性的散点图，其中以性别作为目标属性：每个对象都用自己的类进行了标记，这里使用了不同的形状。图 3.12 显示了不同类的预测属性是如何相互关联的。

需要注意的是，主对角线上方和下方表示的信息相同，这是因为 x 和 y 属性间的相关性与 y 和 x 间的相关性是一样的。由于每个对象的位置根据两个属性的值确定，图 3.12 中显示垂直和水平轴上每个属性的值。矩阵的第 1 行为属性“最高温度”与其他 3 个属性“体重”“身高”和“年数”之间的 Pearson 相关性。类似地，第 2 行显示属性“体重”与“最高温度”“身高”和“年数”之间的 Pearson 相关性。

可以看出，预测属性“身高”和“体重”是正线性相关的，当其中一个的值增加时，另一个的值也随之增加。

我们已经看到主对角线上面和下面表示相同的信息。散点图矩阵的一个版本利用这种

冗余显示每对属性的散点图和相应的相关值，如 Pearson 相关系数。图 3.13 为一个散点图的例子。

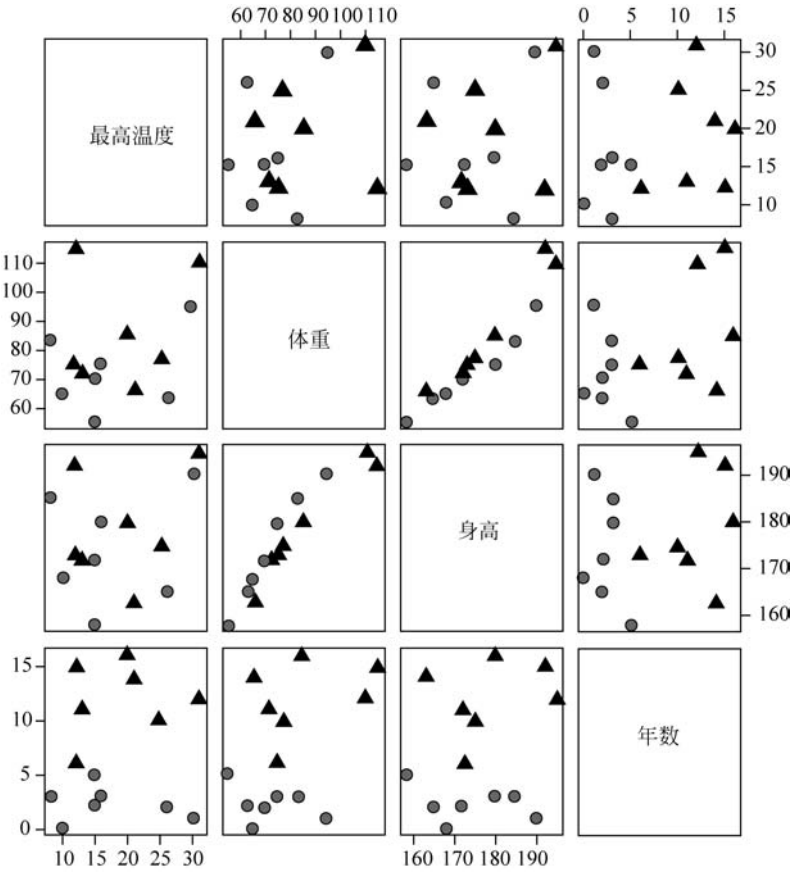

图 3.12　定量属性的散点图矩阵

我们可以使用一个简单的图总结散点图矩阵中的信息。如图 3.14 所示，线性相关矩阵可以被绘制成相关图。图 3.14 中，与两个属性相关的方块颜色越深，它们之间的相关性越强。属性“身高”与“体重”的相关性高，其他属性对的相关性低。正相关和负相关由不同的颜色表示。

和散点图矩阵一样，相关图的主对角线上下也是对称的，因此可以绘制相关值，而不是主对角线上下的彩色方框。

多元数据的另一个常见图是热图，它用方框矩阵表示一个数值表，每个值对应一个方框。矩阵的每行（或每列）都与一种颜色相关联，行（或列）中的不同值由行（或列）的不同颜色表示。热图已广泛应用于生物信息学中基因表达分析。

例 3.9　图 3.15 展示了联系人数据集的简单版的热图。对于垂直轴上的 4 个位置，其中每个都与一个对象相关联，而横轴上的每个位置则与不同的属性相关联。每个属性对应一种颜色，在本例中，对于给定的对象，颜色越深，该对象的属性值越小。

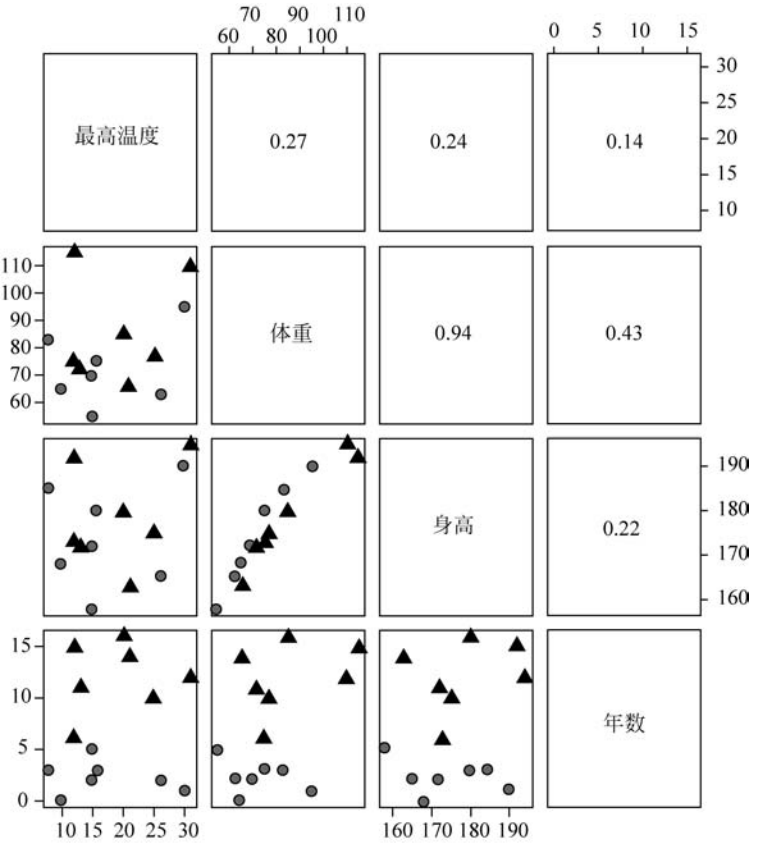

图 3.13　具有额外 Pearson 相关性的定量属性散点图矩阵

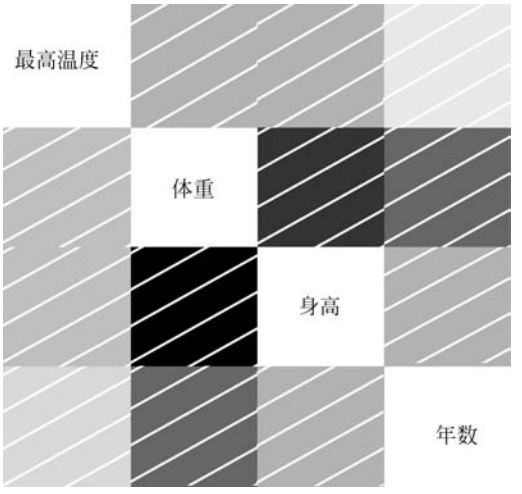

图 3.14　“最高温度”“体重”“身高”以及“年数”
属性间的 Pearson 相关的相关图

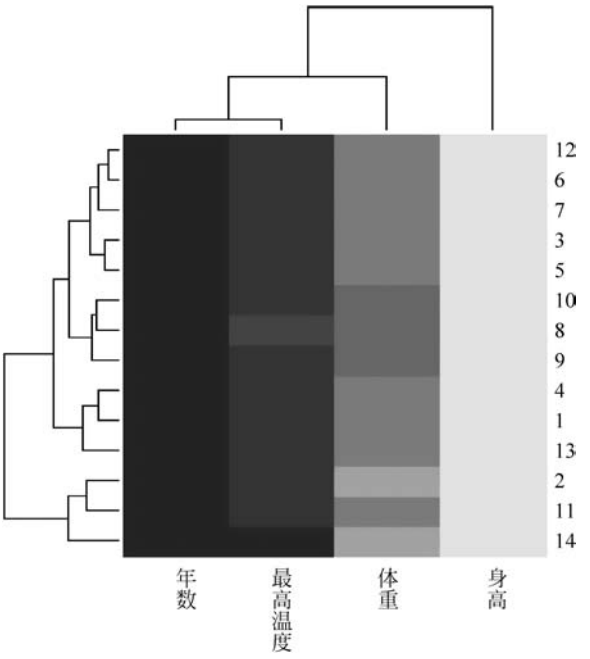

图 3.15　联系人数据集简单版的热图

热图顶部和左侧的图示称为树状图，我们将在第 5 章详细介绍。它们表示根据相互间的相似性进行的属性分组（顶部）和对象分组（左侧）。

我们之前提到过，多元分析图大都是为了定量数据而开发的。由于定性序数属性是直接转换的，所以在图中也很容易使用。随着名义定性数据分析的重要性日益增加，新的图表被创造了出来，其中一个例子是第 2 章介绍的马赛克图，它可以表示最多 3 个定量属性组合的频数。为此，从定性属性中提取了定量值（如频数）。

到目前为止，用于说明数据集所含信息的可视化图例只使用了少量的属性。不过，正如本章开头所提到的，许多实际问题都有几十、几百甚至几千个属性。虽然可以从高维数据集中提取统计度量，但用户要么收到数据中信息的大概，要么无法分析数据，要么被大量信息淹没。

3.4　信息图和词云

3.4.1　信息图

目前，经常使用信息图突出重要事实。理解数据可视化和信息图之间的区别是很重要的。虽然这两种技术都将数据转换成图像，但信息图方法非常主观，是手动生成的，并且要针对特定的数据集进行定制。另外，数据可视化是客观的、自动生成的，可以应用于许多数据集。在本章中，我们已经看到了几个数据可视化的例子，图 3.16 则是一个信息图的实例。

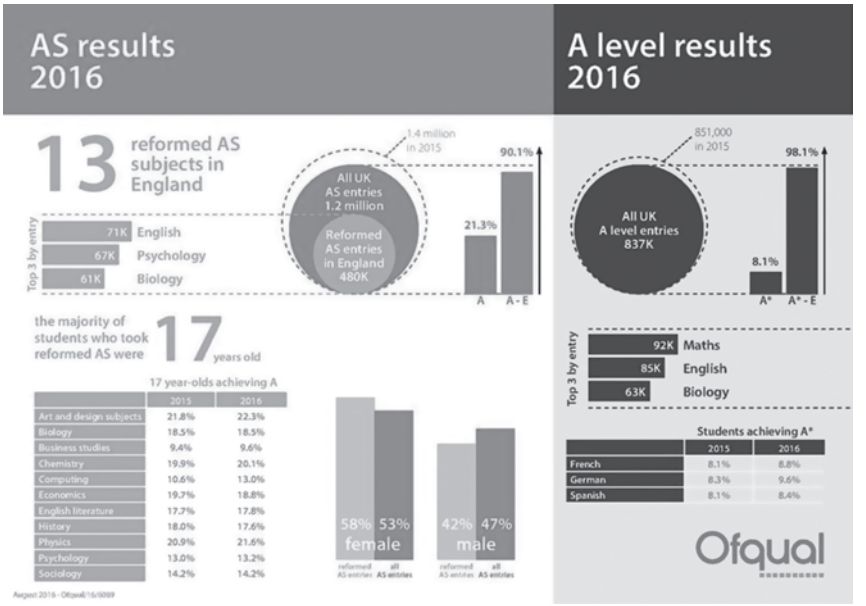

图 3.16　英格兰的资格等级信息图(包含根据开放政府许可证 v3.0 许可的公共部门信息)

3.4.2　词云

词云是一种在文本挖掘中经常用于演示文本数据的可视化工具,它表示给定文本中每个单词出现的频率。一个单词在文本中出现的频率越高,它在其中的大小就越大。由于冠词和介词经常出现在文本中,而数字不是文本,所以在将词云工具应用于文本之前,通常会将它们删除。另一个文本处理操作,也就是词干提取,在使用词云工具之前,也会将文本中的单词替换为它的词干。图 3.17 显示了对一段语句应用词云工具的结果,从中可以看出,在之前的文本中,词干出现频率较高的单词,如 text、word 以及 cloud,用较大的字体表示。

图 3.17　利用词云的文本可视化

3.5 本章小结

本章将单元分析和双元分析扩展到两个以上属性的分析，介绍了频数测量、数据可视化技术和多元分析的统计测量方法。

可以分析多元数据的方法还有很多，而且功能强大。但是，这些方法超出了本书的范围，这里只讨论最常用的技术。其他方法通常出现在更高级的多元分析书籍中，这些方法适用的主题面向的是那些具有初级统计知识的人。

第 4 章将讨论数据集质量的重要性，以及其如何影响接下来的分析步骤，介绍在低质量数据中发现的主要问题以及处理这些问题所需的技术、修改类型所需的操作、规模和数据分布、数据的维数和数据建模之间的关系，以及如何处理高维数据。

3.6 练习

（1）为什么在单元分析和双元分析中使用的一些技术不能用于多元分析？

（2）多元图所提供的信息有何局限？

（3）假如用房屋绘画而不是用 Chernoff 脸谱表示对象，请描述你可以从绘画中得到且用于表示对象的 5 个特征。

（4）平行坐标图的主要问题是什么？如何使这个问题最小化？