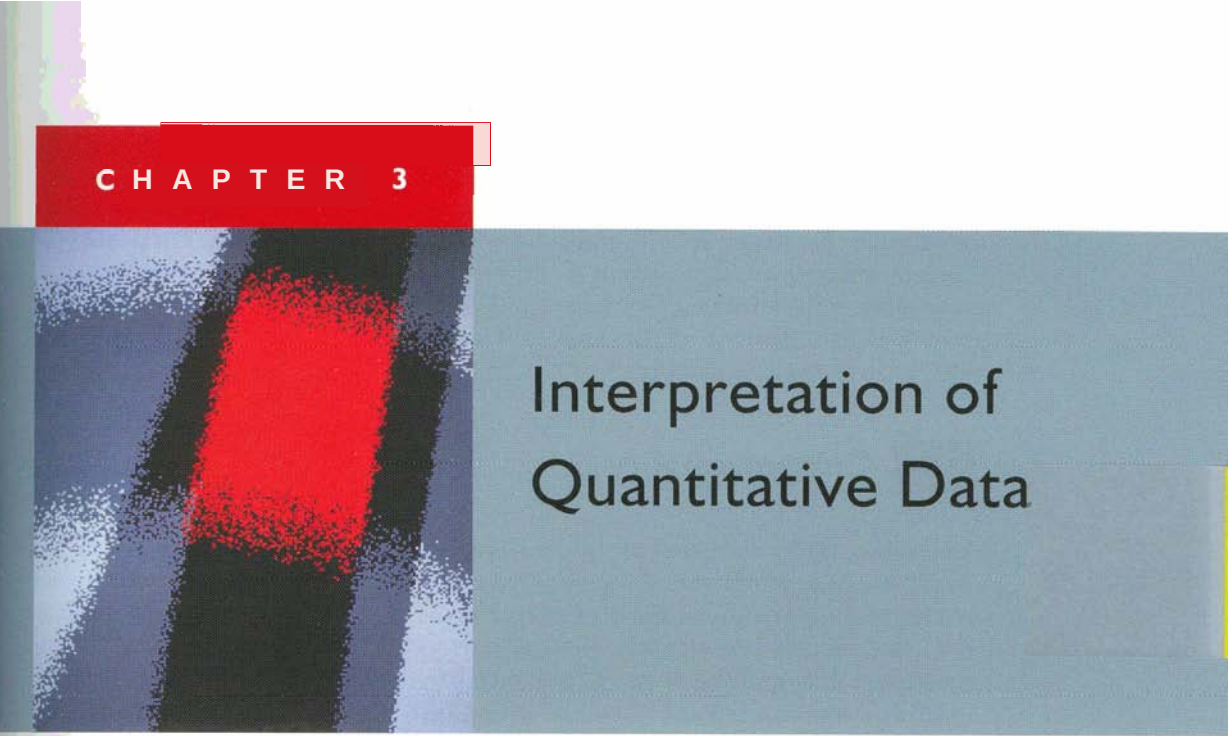




Interpretation of Quantitative Data

3.1 Content

As explained in the Preface, the primary *raison d'être* of this book is the presentation of those advances in information visualization that have been made possible, mainly during the decade from 1990 to 2000, by the emergence of inexpensive powerful computers, high-quality graphics, plentiful memory and responsive interaction. The interpretation of quantitative data by visual means has, of course, been proceeding over a much longer period, as evidenced by the classical works of Sir Ronald Fisher, John Tukey, Bertin and, most recently, Cleveland and McGill. This body of knowledge is certainly relevant to the reader. However, in view of its extent as well as the declared focus and readership of this book, that established knowledge will be selectively summarized here rather than discussed exhaustively. The aim of this chapter is to provide, for a wide readership, an introductory palette of useful techniques, but one which can be extended and underpinned by reference to existing texts according to the reader's interests.¹

¹ Frequently referenced texts include those by Bertin (1981), Cleveland (1993, 1994), and Cleveland and McGill (1988). The reader is also referred to Card *et al.* (1999) for more recent comment.

3.2 Overview

An overview of this chapter can be obtained by considering the illustrative task of selecting one object from many on the basis of the numerical values of their attributes. An **example** is the purchase of a new car: here the attributes will include price, passenger capacity, petrol consumption and other quantitative measures.² The problem faced by the designer of an appropriate visualization tool is that of presenting the data, and allowing it to be rearranged, in such a way as to facilitate a decision.

3.3 Dimensionality

The difficulty of designing an interactive display is strongly influenced by the number of attributes – more generally referred to as variables – involved, and sometimes referred to as the dimensionality of the problem. If the cars are characterized by only one attribute (for example, cost) we are said to be dealing with univariate data and the design of the visualization tool is relatively – though not entirely – straightforward. If two attributes are involved we are said to be handling bivariate data, and the task of the visualization tool designer is a little more challenging. It is at this point that one is reminded of the African tribe whose language has words for 'one' and 'two', but only their word for 'many' for any number greater than two, because the task of enhancing a user's understanding of trivariate data (three attributes) suddenly becomes much more taxing. Indeed, in view of the many and varied situations in which more than three attributes are involved (hypervariate data) it is usually preferable, for trivariate data, to address the problem posed by many attributes where, as for that African tribe, 'many' is any number greater than two.

In what follows we shall, in fact, follow the taxonomy just introduced and consider, in turn, univariate, bivariate, trivariate and hypervariate data.³

3.4 Univariate data

Let us suppose there are 30 cars associated with some selection we have to make, and that they are characterized by the numerical value of just one attribute, that of price; here we have an example of a collection of measurements or attributes of a single quantitative variable. The details could be presented in a table (Table 3.1) or, almost certainly more effectively, in the form of points against some scale (Figure 3.1(a)). Space considerations will determine whether the corresponding objects can be labeled to enhance comprehension

² The fact that, in real life, the final choice may be heavily influenced by non-quantitative features such as colour and shape is not, as we shall see, a trivial consideration (see Chapter 5).

³ This is, in fact, the taxonomy and pedagogical approach used by Cleveland (1993).

TABLE 3.1 The price and make of a collection of cars

<i>Car</i>	<i>Price (£)</i>
BMW	51,395
Mercedes	50,850
Mercedes	41,000
Saab	39,085
BMW	38,000
Ford	36,950
BMW	34,550
Morgan	32,000
Rover	31,300
Ford	29,250
Saab	28,750
Mercedes	28,000
Ford	27,600
Ford	25,950
Land Rover	24,000
Jeep	23,200
Nissan	20,000
Vauxhall	18,500
Nissan	17,400
Ford	17,000
Ford	16,500
Nissan	15,500
Ford	14,950
Ford	12,000

(Figure 3.1(b)). Both examples in Figure 3.1 have useful features; the lowest and highest values are easily noted, the general distribution is clear and any 'bunching' is obvious.

Interpretation may be eased by introducing various forms of aggregation, for example by indicating the average of all the prices. Particularly valuable are Tukey Box Plots (Figure 3.2) in which the box indicates the 25th, 50th and 75th percentiles, the bars the 10th and 90th percentiles, and where any 'outliers' beyond the latter two percentiles are shown individually. An alternative is to retain all the individual points, and simply superimpose indications of the same percentiles (Figure 3.3). Other presentations of the same data, and involving aggregation, are possible (Tufte, 1983), and one is illustrated in Figure 3.4.

FIGURE 3.1

(a) Univariate data plotted against a scale
(b) Labeling of the data points may be difficult

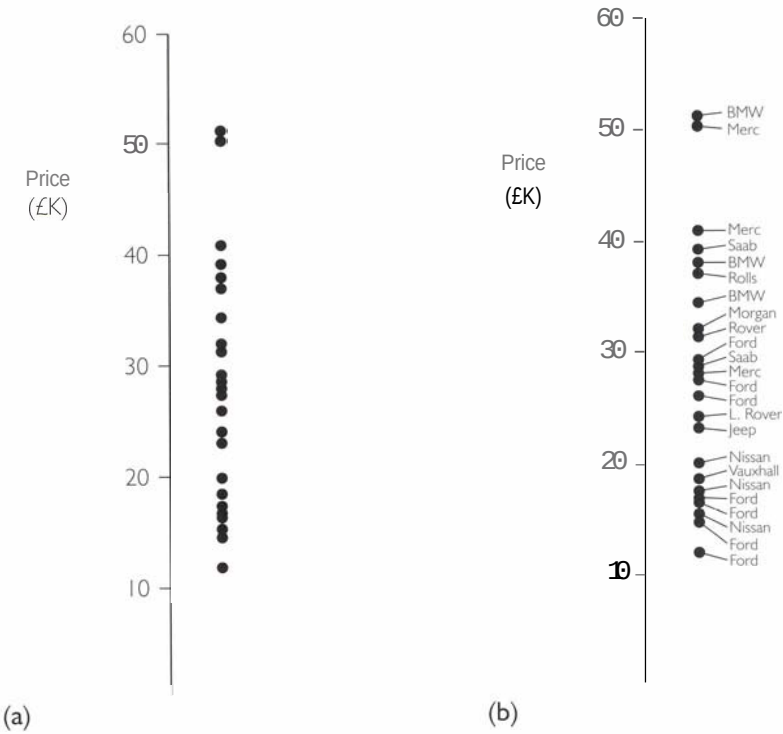
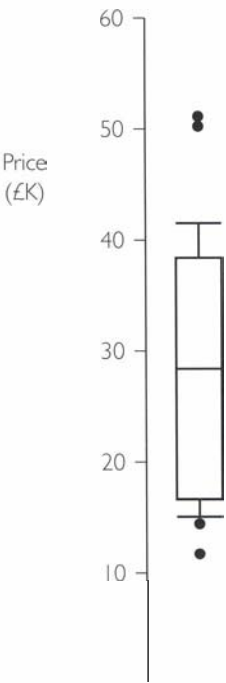


FIGURE 3.2

Tukey Box Plot representation of univariate data



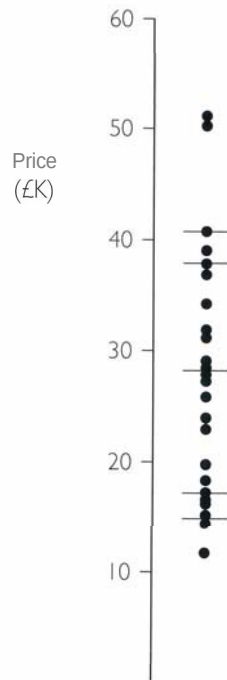


FIGURE 3.3
An alternative
representation
of key
percentiles

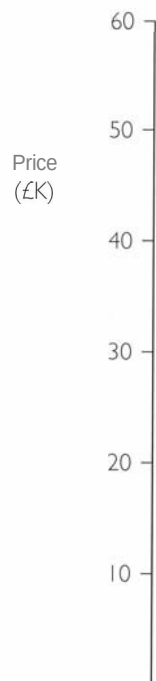


FIGURE 3.4
An alternative
representation
of the Tukey
Box Plot

FIGURE 3.1

(a) Univariate data plotted against a scale
(b) Labeling of the data points may be difficult

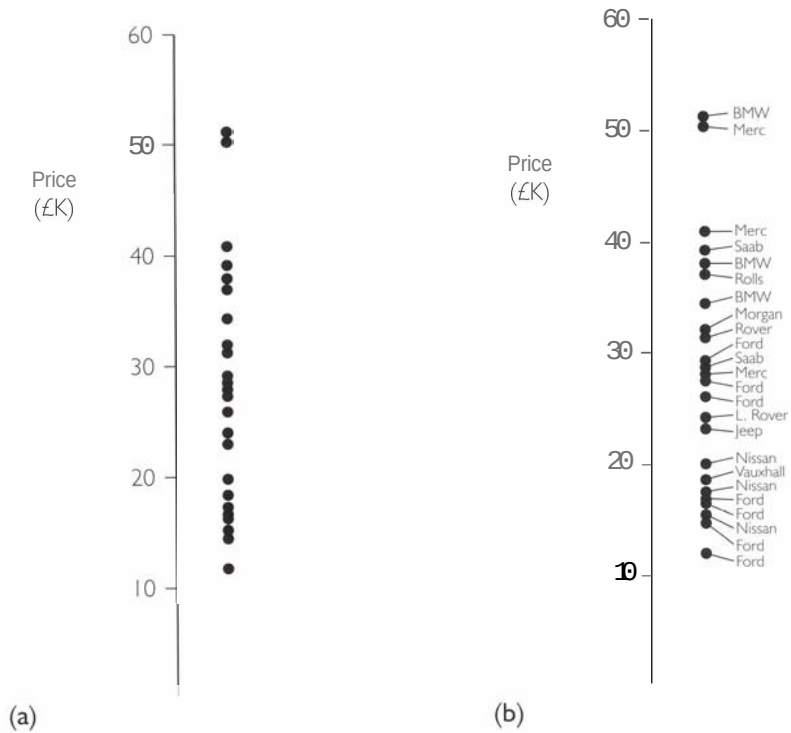
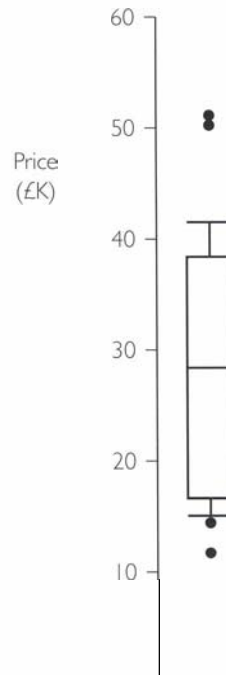


FIGURE 3.2

Tukey Box Plot representation of univariate data



The number of data points can be – and usually is – quite large, in which case the ability to zoom to examine selected ranges in more detail can be **valuable** (Figure 3.7). Simple zooming, however, only allows a user to see bigger blobs: a much more effective operation is logical zooming, wherein more and more detail is revealed as magnification increases (Figure 3.8), an activity we shall later call semantic zooming.

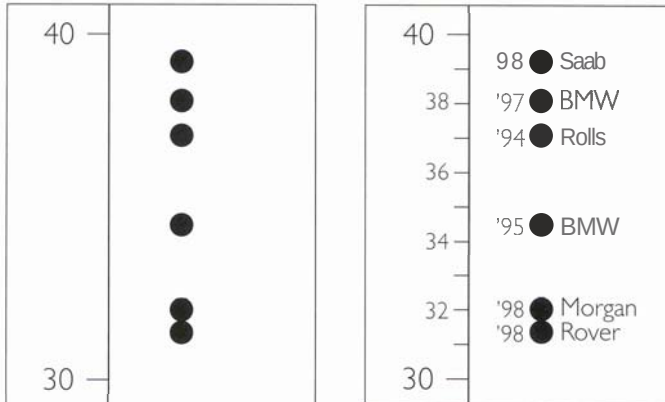


FIGURE 3.7
Result of
zooming into
univariate data

FIGURE 3.8
Result of a
logical zoom

3.5 Bivariate data

A conventional approach to the visualization and interpretation of bivariate data is typified by a two-dimensional plot (Figure 3.9) of one variable against the other, a presentation called a scatterplot. In the house-hunting **example** shown a user can clearly identify global trends (Price generally increases with Number of Bedrooms) and local features such as trade-offs, **as well** as outliers that could **prove** both unexpected and worth further investigation (such as the **five-bedroom** house selling for £50,000)

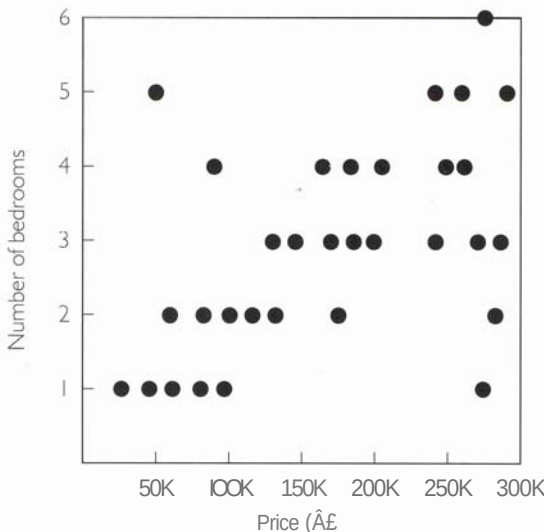


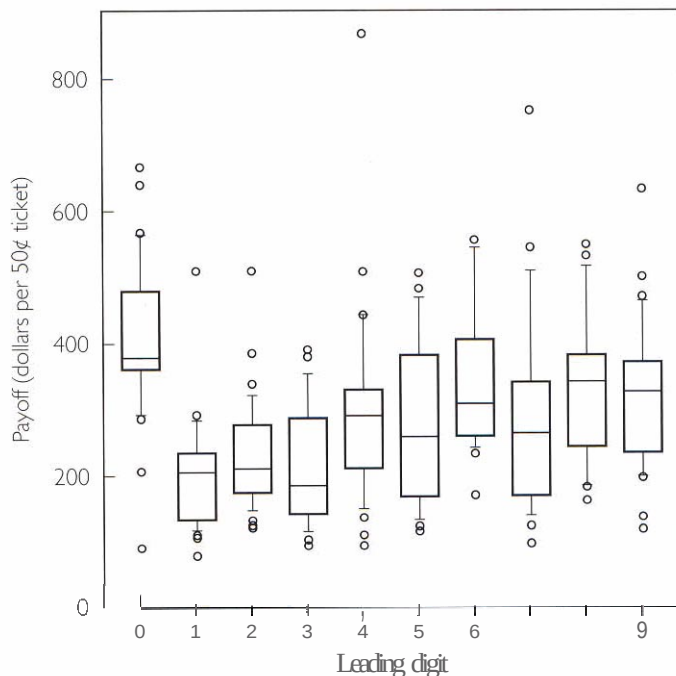
FIGURE 3.9
A scatterplot

In situations in which one of the two variables is controllable, and the data is thereby grouped, the box plot approach can again assist interpretation. In the example of Figure 3.10 one of the variables is the leading digit (of three) selected by a player in the New Jersey lottery, and the vertical axis is the payoff (Becker and Chambers, 1984). It is clear, for example, that the payoffs for three-digit numbers starting with a zero tend to be high, which means that those who bet avoid them. The plot also suggests that odd first digits are preferred to even ones.

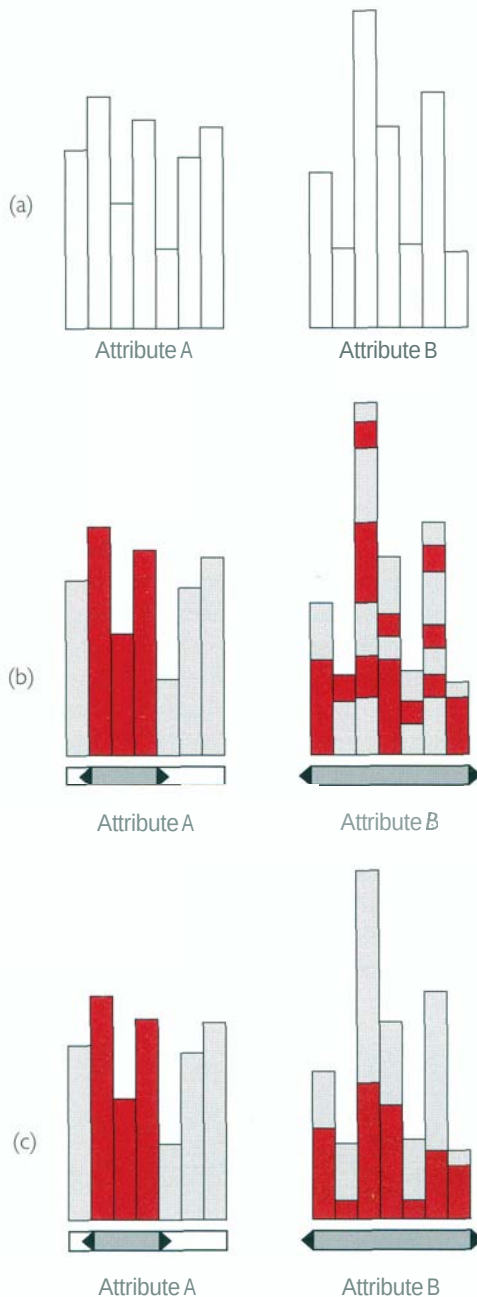
FIGURE 3.10

Box plots used to represent bivariate data

Source: Redrawn from Cleveland (1985)



An approach to the display of bivariate data, and one which is extensively exploited later in the book (especially for hypervariate data) involves the use of histograms. While a separate histogram (Figure 3.11(a)) associated with each variable is of limited use as it stands, two enhancements can lead to an extremely useful visualization tool, and one whose value is enhanced by its application to hypervariate data. We first provide, below each histogram, a scale with adjustable **limits** by means of which a subset of the data can be identified (Figure 3.11(b)). Secondly, we ensure that the objects so identified by both sets of limits are colour-coded (Figure 3.11(b)). The display now becomes much more valuable since the influence of one variable on the other can be manually explored. An apparently minor modification in which selected (red) objects are 'ranged down' (Figure 3.11(c)) considerably enhances the value of the linked histograms, an advantage that can readily be appreciated when used dynamically. It is not difficult, as we shall see in Chapter 5, to arrange additionally for the manual adjustment of a **range** of variable A up and down the scale (that is, the limits move together), allowing a global impression of the relation

**FIGURE 3.11**

(a) Histograms of two attributes
 (b) Histograms with adjustable limits for selection purposes
 (c) Selected objects are 'ranged down' for easier interpretation

between A and B to be gained. The enormous value of dynamically adjusted selection emphasizes the differences between visualization tools that are passive (for example, Minard's map), active (where a single action causes rearrangement) and **dynamic**, where continuous movement of a control causes **continuous** rearrangement.

3.6 Trivariate data

Living as we do in a three-dimensional world one would imagine that a three-dimensional display of data would be regarded as 'natural' in some sense and therefore ideal for the display of data. While that might be true to some extent in the (usually expensive) context of virtual reality, the problem arises with the fundamentally two-dimensional representation of a three-dimensional space. Figure 3.12, for example, provides a two-dimensional view of a three-dimensional space in which four houses have been placed according to their numerical values of three attributes, Price, number of Bedrooms and Time to reach one's

FIGURE 3.12
A 2D
presentation of
a 3D plot



workplace. But can we decide if A has a greater value of Price than B? No, we can't. To answer this and similar questions we could project the points onto all pairs of axes (Figure 3.13). More usefully, however, we can **rearrange** these plots in an organized manner to form a scatterplot matrix (Figure 3.14).

FIGURE 3.13
Projection of
the points in
Figure 3.12
onto all pairs
of axes



At first sight the value of a scatterplot matrix is not apparent. Since the matrix contains as **many** scatterplots as there are pairs of parameters it can become a little **unwieldy** for more **than** about five parameters. Moreover, we **are** faced with the drawback that for **N** objects with three attributes we now have a collection of $3N$ points; if there are 4 attributes the number of points **increases**

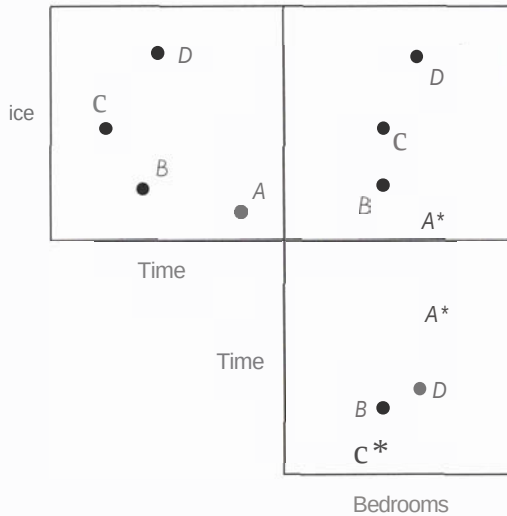


FIGURE 3.14
A scatterplot matrix for the 3D plot of Figure 3.12

to $6N$ while for six variables we have $15N$. Additionally, the labeling of points can often be impossible in view of the number of points and the screen space available. Fortunately the technique of brushing can aid interpretation. For the scatterplot matrix of Figure 3.14 it is possible to identify a group of points in one of the scatterplots (Figure 3.15) whereupon those objects are highlighted in some way in all the other scatterplots. In this way the effect of one parameter on the relation between the other two can readily be explored. Brushing is particularly useful for hypervariate data, as we shall see later in the book. It is a particular example of subsetting (that is the selection of subsets) which finds wide application. Recently the brushing concept, illustrated here in the context of data, has been extended to the brushing of structure (Fua *et al.*, 1999).

The difficulty associated with the interpretation of Figure 3.12 is also experienced with surfaces such as the example of Figure 3.16. What, for example, is

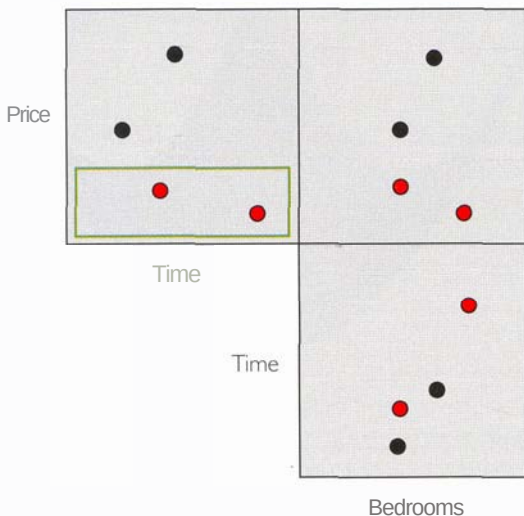
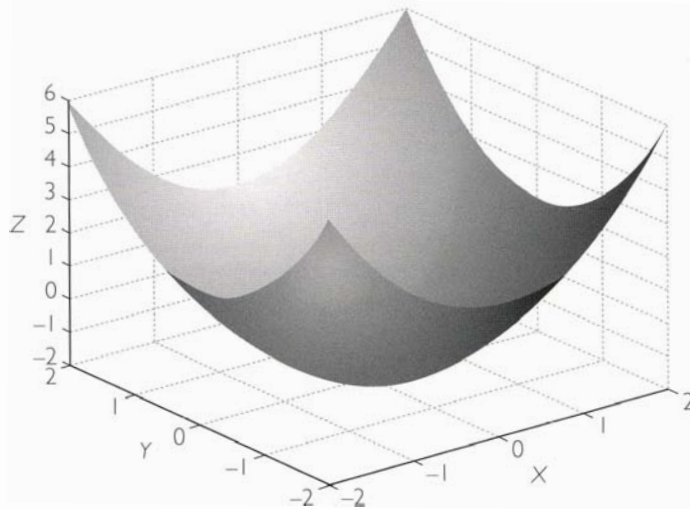


FIGURE 3.15
Brushed points are highlighted on all plots

FIGURE 3.16

A surface in 3D space

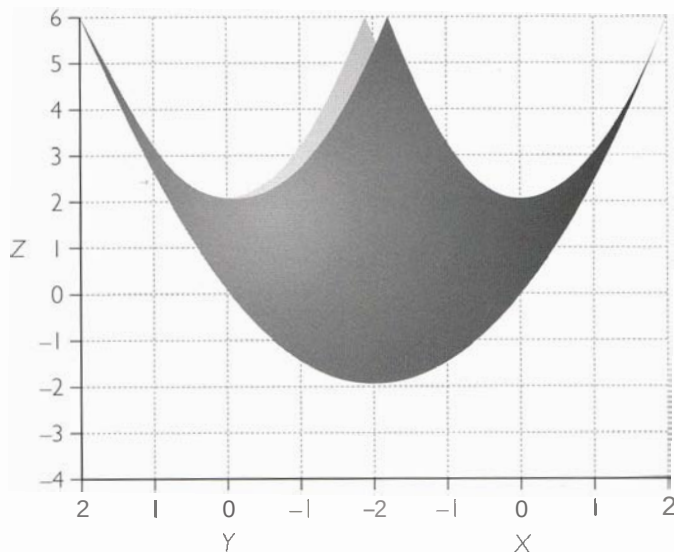


the minimum value of Z ? An answer to this and similar questions can be obtained if the viewpoint can be adjusted (Figure 3.17) and if we can use techniques such as 'flooding' (Figure 3.18). A more general approach, reflecting the comment (relevant to Figures 3.12 and 3.16) that 'for 3D to be useful, you've got to move it', is embodied in the **spinplot** (Fisherkeller et al., 1974). With this tool, the **view** of a three-dimensional arrangement of points, or a surface, can be freely, easily and continuously adjusted by a user (Figure 3.19(a) and (b)).

As remarked earlier, when the number of dimensions is three it may be appropriate to consider techniques that have been invented to support the visualization of hypervariate data, a topic discussed in the following section. Also, for any number of dimensions, techniques such as animation, described in Chapter 2 in the context of sensitivity circles, can often be exploited to 'add another dimension'. If they are, however, one must be aware of any danger asso-

FIGURE 3.1

A view of the 3D space of Figure 3.16 useful for finding the minimum value of Z



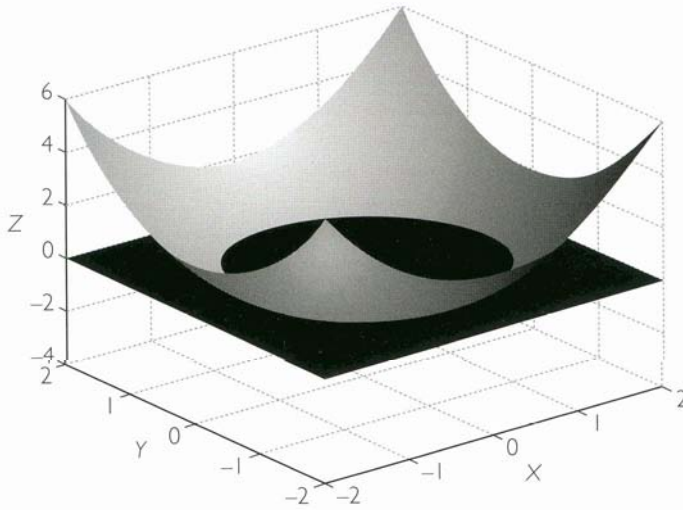
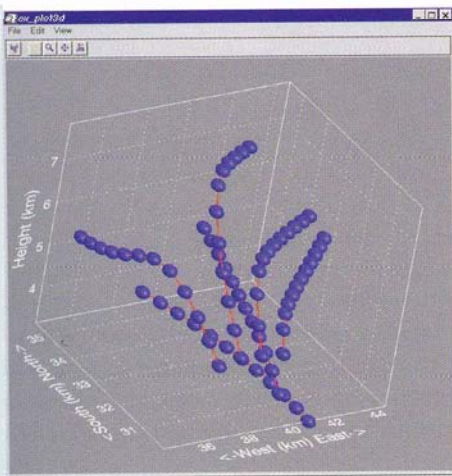
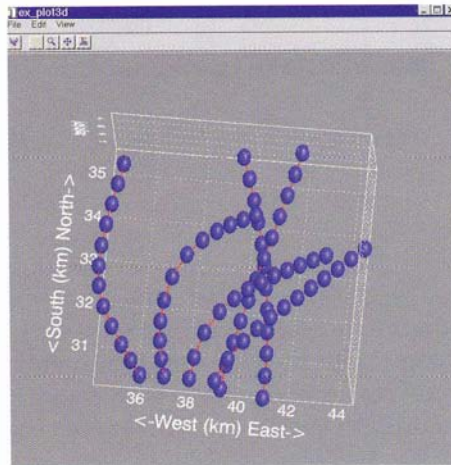


FIGURE 3.18
The technique
of 'flooding'



(a)



(b)

FIGURE 3.1
A spinplot
allows viewing
in any direction
Source: Images
produced by *IDL*, and
reproduced by
permission

ciated with the fact that all dimensions may not be given equal perceptual weight: in some applications (see the Influence Explorer, p.160 in Chapter 9) it is important that the display should not be biased to favor the user's view, interpretation or control of a particular variable.

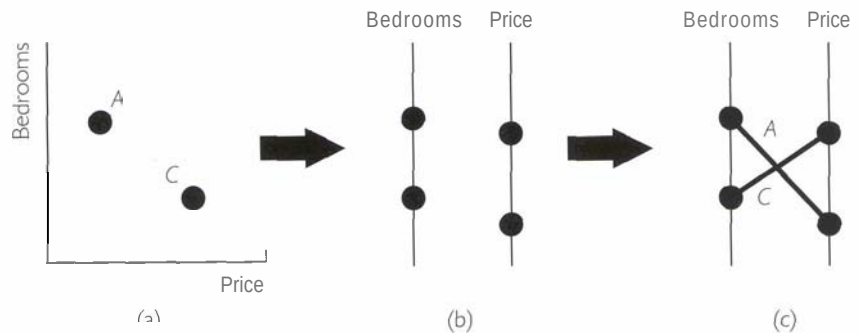
Hypervariate data

The challenge of visualizing hypervariate (also termed multivariate) data is ongoing, and certainly motivated by the many situations in which the inter-relationships between many variables are of vital interest; engineering design, house-hunting and financial design are just a few of many available examples.

Parallel coordinate plots (Inselberg, 1985, 1998; Wegman, 1990) offer a tried and tested technique which can handle a large number of variables. The principle is simple: it is that of taking all the axes of the multidimensional space and arranging them in order, but parallel to each other. The principle is illustrated in Figure 3.20 with the simple example of two attributes (and hence two parallel coordinates) and two houses (A and C of Figure 3.13). Essentially, we use that earlier form of presentation (Figure 3.20(a)) but then take the two axes **Price** and Bedrooms apart and draw them in parallel (Figure 3.20(b)): each house now appears once on each axis. As it stands, of course, there is nothing to identify which house is which – do we have a cheap house with few bedrooms and an expensive one with more, or an inexpensive house having more bedrooms than the more expensive one? The ambiguity is simply resolved by joining corresponding entries on each scale, as shown in Figure 3.20(c). The two axes are

FIGURE 3.20

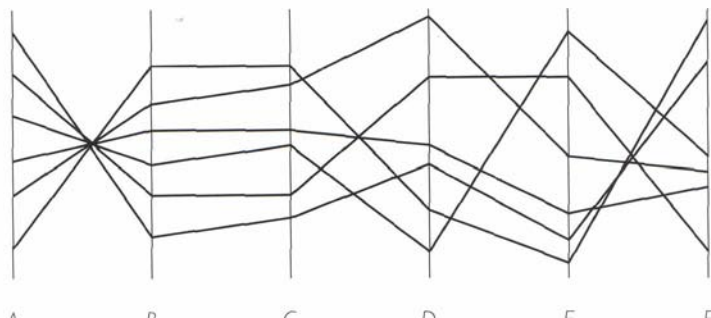
The principle of parallel coordinate plots

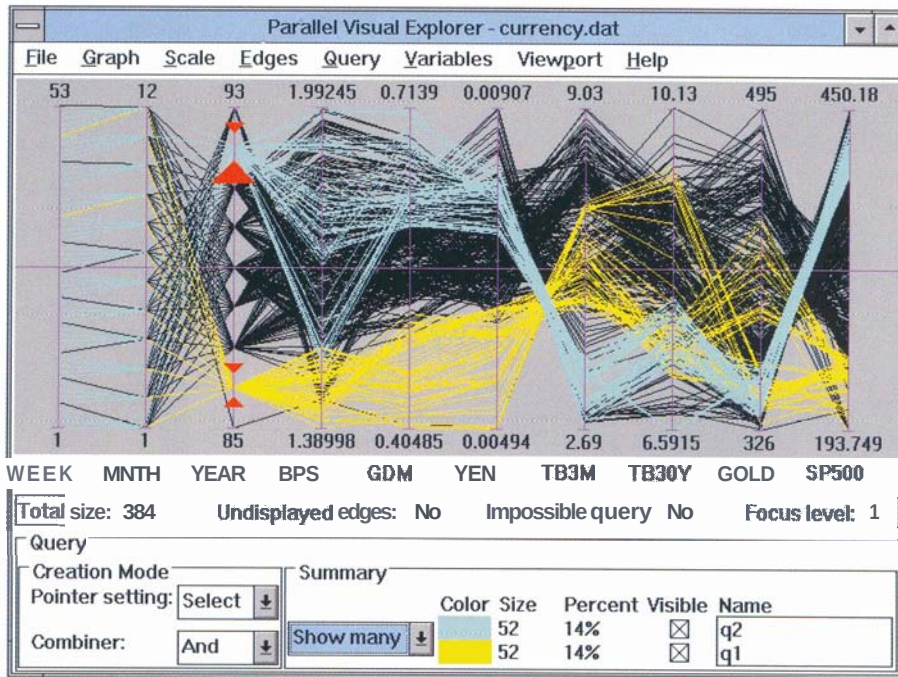


referred to as parallel coordinates and the entire presentation as a **parallel coordinate plot**. By reason of its very construction, the concept of parallel coordinates can be extended to any number of variables, as illustrated in Figure 3.21. An example of a parallel coordinate plot relevant to the financial domain is shown in Figure 3.22, where **color** is used as an additional coding technique. As well as handling quantitative data, parallel coordinate plots can also handle categorical data.

FIGURE 3.21

The principle of parallel coordinate plots applied to six variables



**FIGURE 3.22**

A parallel coordinate plot for financial variables

Source: Reproduced by permission from the IBM Corporation

Certain relationships between variables can easily be deduced from the appearance of the plot. In the simple example of Figure 3.21 there is clearly a trade-off between variables A and B, and a strong correlation between B and C, though it has to be said that ease of interpretation can be strongly influenced by the ordering of the axes. However, experts in the interpretation of parallel coordinate plots can derive a great deal of understanding from these plots. Inselberg (1998), for example, provides an interesting illustration related to the enhancement of the manufacturing yield of a process making silicon chips, showing the considerable amount of useful information that can be deduced. More recently, Inselberg and Avidan (1999) have introduced automation to facilitate the elicitation of knowledge from parallel coordinate plots. A significant advantage of parallel coordinate plots, shared by some other visualization tools, is that all variables are treated equally.

The mosaic plot, used in the previous chapter to illustrate the potential of rearrangement, is a useful technique for handling high-dimensional data. If we were to return to the example in Chapter 2 and imagine that the person recording the hair and eye colours of the 592 students also recorded gender, the mosaic plot of Figure 2.8, repeated here as Figure 3.23(a), could be modified as shown in Figure 3.23(b). An extension to even higher dimensionality is illustrated by the Titanic disaster of April 1912, when 1731 of the 2201 passengers and crew were lost (Dawson, 1995). Table 3.2 shows the raw data, involving four variables: Gender, Survival, Class and Adult/Child. A number of mosaic plots can be generated (Friendly, 1994), and appropriately interpreted: that shown in Figure 3.24 involves all four attributes.

FIGURE 3.23
(a) The Mosaic Plot of Figure 2.9 for hair and eye colours
(b) Modification of the Mosaic Plot to include gender data

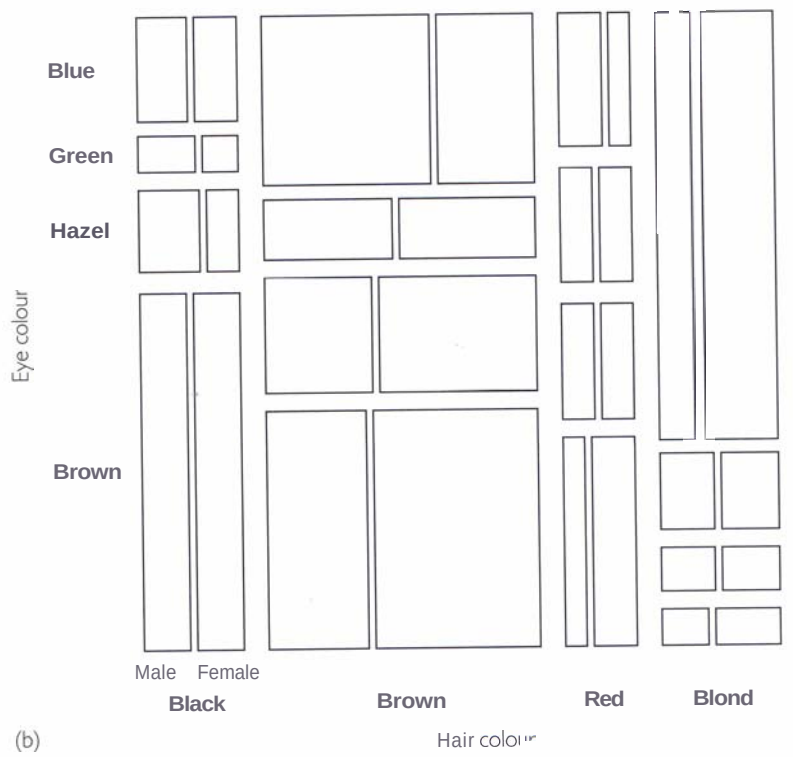
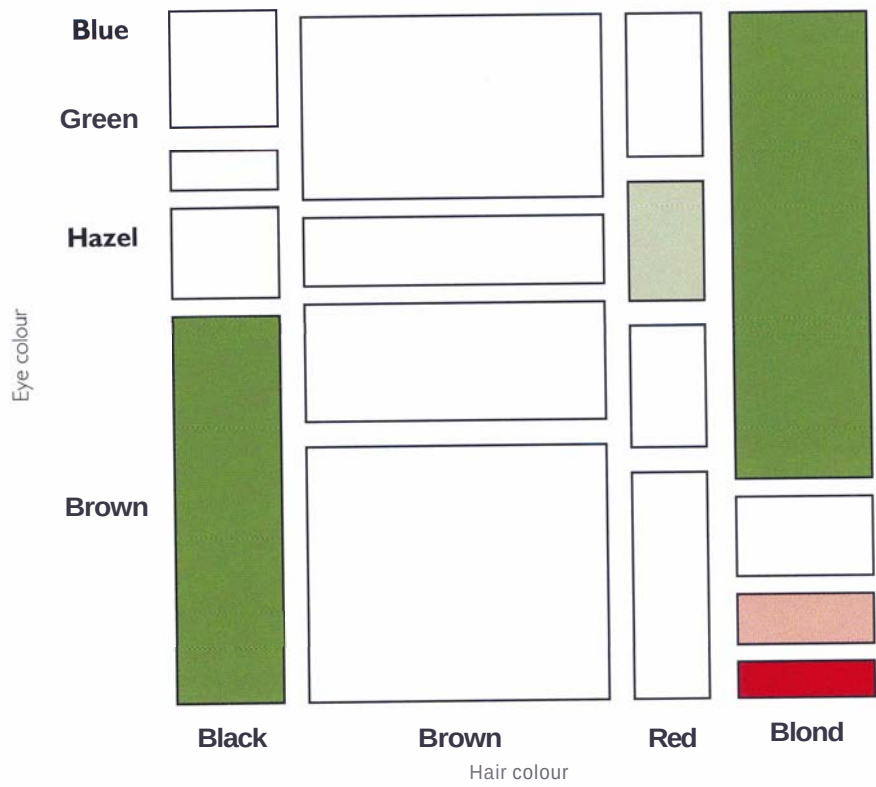


TABLE 3.2 Data concerning the Titanic disaster

Survived	Age	Gender	Class			
			1st	2nd	3rd	Crew
No	Adult	Male	118	154	387	670
Yes			57	14	75	192
No	Child	Female	0	0	35	0
Yes			5	11	13	0
No	Adult	Male	4	13	89	3
Yes			140	80	76	20
No	Child	Female	0	0	17	0
Yes			1	13	14	0

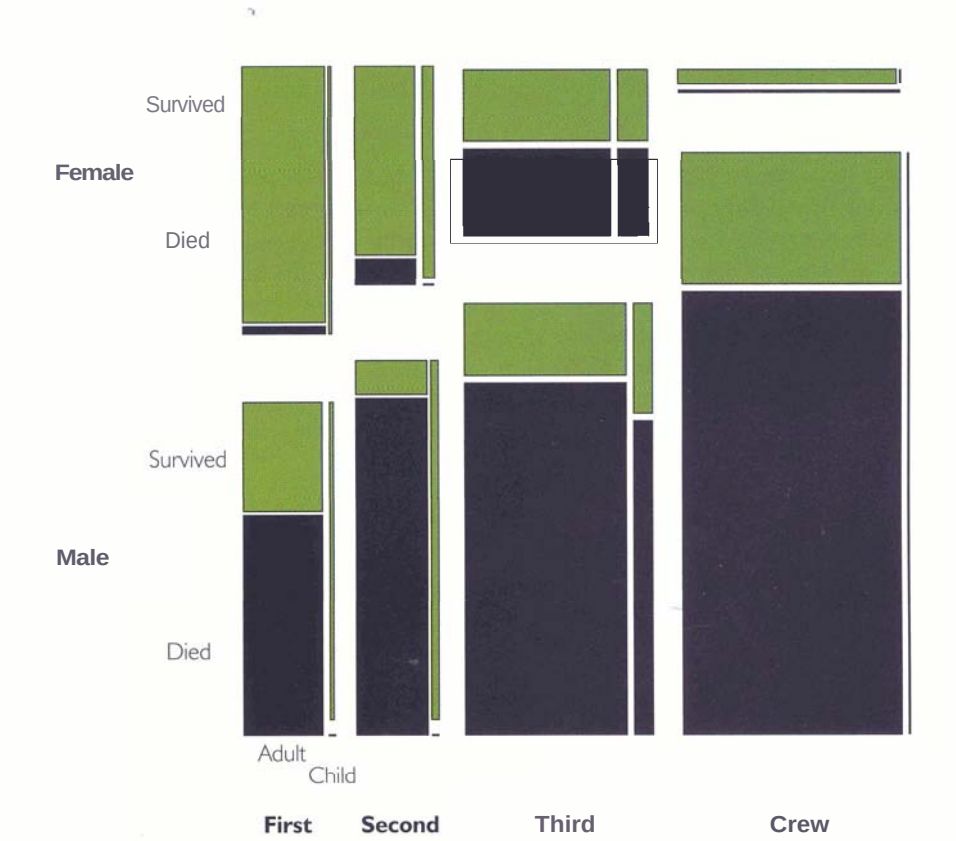
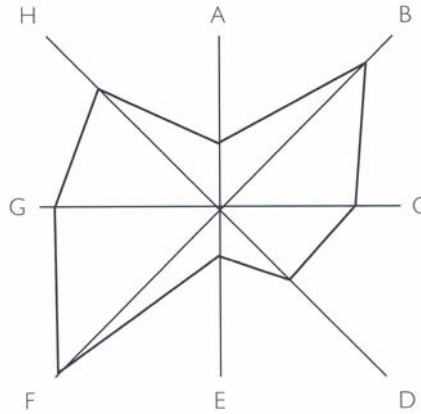


FIGURE 3.24
A Mosaic Plot
for the Titanic
data of Table
3.2

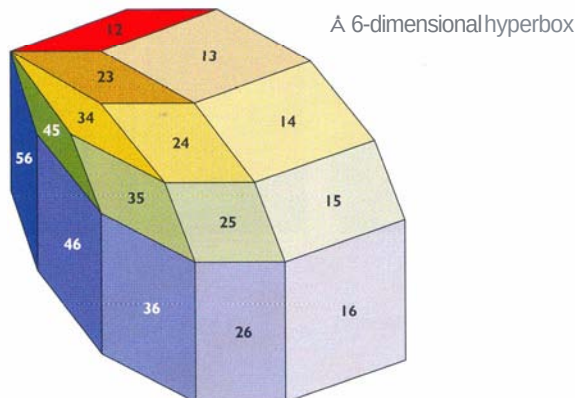
A longstanding means of presenting hypervariate data, and which has much in common with the parallel coordinate plot, is the **starplot** in which the axes radiate in a star shape from a common origin (Figure 3.25). Thus, a number of objects, each represented by a star plot, might be compared, at least qualitatively, on the basis of their shapes. Alternatively, as with the parallel coordinates, a number of objects can be represented on the same star plot.

FIGURE 3.25
A starplot for
eight variables



The **hyperbox** (Alpern and Carter, 1991) has the appearance (Figure 3.26) of a solid three-dimensional object, but is so constructed that all possible pairs of variables are shown plotted against each other. Conceptually there is no limit to the number of dimensions that can be handled. One useful feature, which avoids the ordering problem associated with parallel coordinate plots, is that all combinations of two variables are shown. With modern technology any pair can be brought to the front with Cartesian axes, with all others still visible and providing useful context: brushing facilities could easily be added. The hyperbox has been used to observe the behavior of a computer algorithm.

FIGURE 3.26
A hyperbox
handling six
variables



3.8 Comment

We have made no comment concerning the intrinsic or relative effectiveness of the various techniques for presenting quantitative data for two reasons. First, because the value of a particular technique depends very much upon the application domain, the user and the task being performed. Second, because a combination of techniques, drawing from this chapter as well as the next, is typically involved in a real world application, and we currently have no way of evaluating the relative merits of techniques combined within a real and usually complex visualization tool of the sort encountered in Chapters 5, 7, 8, 9 and 10. It is for these reasons that we concentrate, in this chapter and the next, on providing a rich palette of techniques from which the visualization tool designer can draw on the basis of judgement and experience.