A second scenario in which the awareness of a single number is crucial is a surgical operation and particularly a major one lasting many hours. Any change in vital human signs such as blood pressure is a cause for concern, especially for the anaesthetist (as well as the patient), so it is essential that any change in such a vital sign be noticed immediately. The anaesthetist is, of course, presented with comprehensive visual information in the manner illustrated in Figure 3.10, but the effort of paying careful attention to such a presentation for many hours would be considerable. The simple solution is to encode vital signs in sound (Watson *et al.*, 1999; Watson and Sanderson, 2004): whatever the anaesthetist is attending to (perhaps on the telephone to respond to an urgent enquiry from a ward), he or she will immediately be aware of a change in the pitch or repetition rate of a constant sequence of 'beeps'. Another reason for introducing this example is that it emphasizes the fact that data can be encoded in ways other than visual and that the 'visual' in visualization should not be misinterpreted.

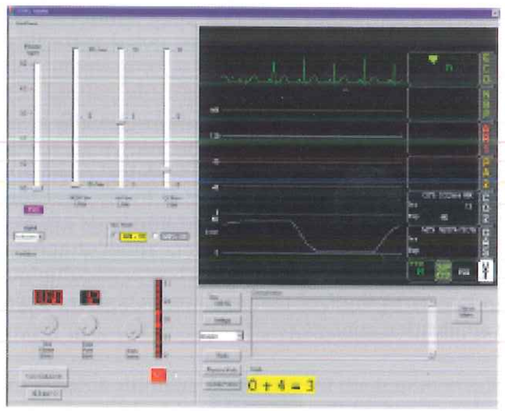


FIGURE 3.10
Representations of the vital signs of a patient during an operation. The difficulty of paying constant attention to such a display throughout a long operation has led to the encoding of vital signs in the pitch of a frequently repeated 'beep'. A change in pitch is immediately apparent wherever the gaze of the anaesthetist is directed
(Image by kind permission of Marcus Watson)

A collection of numbers

Notwithstanding the importance of representing a single number, as in an altimeter, a more common situation is one in which univariate data about a *number* of objects is of interest. Not surprisingly, well-established techniques are available (Cleveland, 1994; Tufte, 1983), though the field of information visualization is such that new ones continue to be invented.

Price data for a number of cars can, for example, be represented as dots on a linear scale (Figure 3.11). But how effective is this representation? A quick overview shows the general distribution of car prices; a more detailed examination will estimate the average price accurately enough to judge 'affordability'; further examination will disclose maximum and minimum prices. A valid question, therefore, for any representation is, 'How will the representation be used?' Many of the *initial* questions in the mind of a car buyer would be answered by a Tukey Box Plot of the same data (Figure 3.12). The box contains the *median* price (the middle line, which divides the number of data items into halves). The two ends of

the box represent the 25 percentile (below which one quarter of the data items are to be found) and the 75 percentile. The 5 and 95 percentiles are indicated by the horizontal bars and the outliers are retained as points. Here, much of the data is being *aggregated*, to reflect the fact that precise detail is often not needed; we are now representing *derived values*.

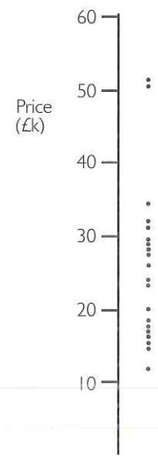


FIGURE 3.11
Each dot represents the price of a car

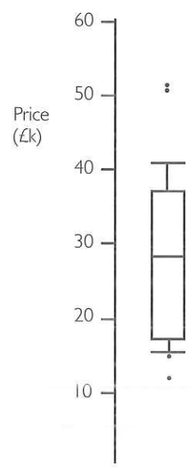


FIGURE 3.12
A Tukey Box Plot of the data represented in Figure 3.11

Another familiar representation of univariate data is the *histogram* (Figure 3.13). Much has been written about the need for appropriate choice of bin sizes and will not be repeated here. The essential point to note is that we are again representing *aggregate* properties – or derived values – of the data in a manner that can support both 'at a glance' awareness and the need for more precise understanding. The histogram is only one of many useful representations of numerical data that are more concerned with derived values. For example, if we 'push over' the columns of the histogram and join them together we obtain the *bargram* (Figure 3.14), already familiar from Chapter 2. While the relative count in a bin is now reflected in the bin's width, certain characteristics of the data are lost or not obvious at first sight – for example, the existence of outliers and the emptiness of a particular range. This omission may be immaterial in certain applications.

Values need not be numerical – they can be *categorical* or *ordinal*. The makes of cars can include Ford, Nissan, Toyota, Ferrari, etc., so that a bargram appearing in an online car sales display like the EZChooser (Wittenburg *et al.*, 2001) might contain this univariate categorical data as shown in Figure 3.15. Ordinal data can be found in the sales volume of a shop for the various ordered days of the week (Figure 3.16).

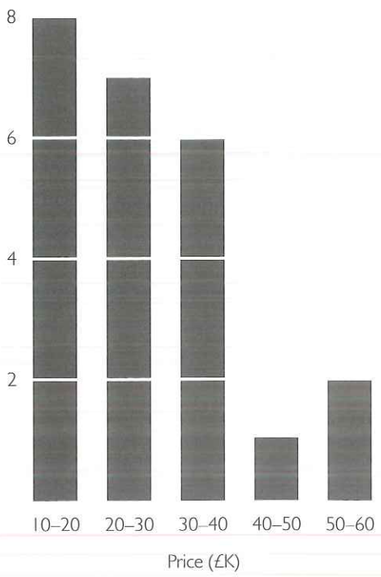


FIGURE 3.13 A histogram of univariate data

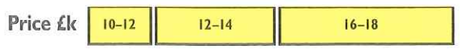
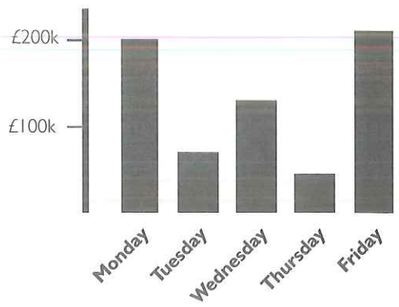


FIGURE 3.14 A bargram representation of univariate data, obtained by 'tipping over' the columns (bars) of a histogram and joining them end to end, ignoring any null bins

FIGURE 3.15
A bargram representation of univariate categorical data

FIGURE 3.16
A histogram of univariate ordinal data.



3.1.2 Bivariate data

A conventional approach to the representation of bivariate data is the *scatterplot*. For a collection of houses, all characterized by a price and the number of bedrooms, each house is represented by a point in two-dimensional space with

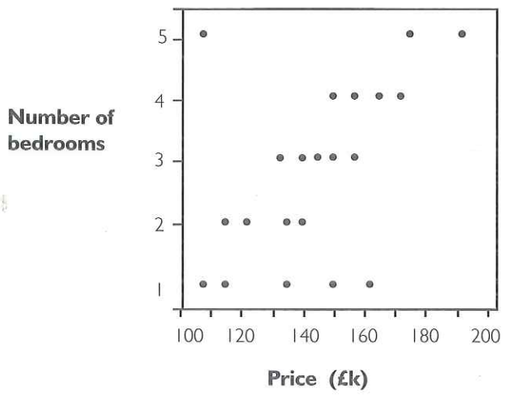


FIGURE 3.17
A scatterplot of bivariate data. Each point indicates the price and number of bedrooms associated with a house

axes associated with these two attributes (Figure 3.17). This representation affords an awareness of a general trend (more money, more bedrooms), of local trade-offs (less money, more bedrooms) and of outliers that may be interesting and might not have been anticipated (a five-bed house for £110,000). Any conventional query system requiring a precise formulation of housing needs would not encourage the specification of such an unanticipated result and is one of the many reasons why information visualization can be so valuable.

A special case of the scatterplot is a time series, in which one axis represents time and the other some function of time. The importance of time-varying data, for example in medical and climate studies, is such that many representation techniques as well as visualization tools have been developed to allow understanding to be derived from a time-series plot. The performance of one time-series query tool is illustrated in Figure 3.18 in the context of a data set containing 52 weekly stock prices for 1,430 stocks (Hochheiser and Shneiderman, 2004). The graph overview of Figure 3.18(a) shows the entire data set, providing some idea of density and distributions. In Figure 3.18(b) a single timebox limits the display to those items with prices between $70 and $250 during days 1 to 4. Subsequent queries add additional constraints, selecting items that have prices between $70 and $95 during days 7 to 12 (Figure 3.18(c)) and for prices between $90 and $115 for days 15 to 18 (Figure 3.18(d)).

An alternative representation of a time series (Figure 3.19), perhaps more suited for gaining an initial impression of data, is illustrated by the level of ozone concentration above Los Angeles, each square associated with one day and coloured to indicate the ozone level. The figure, which represents ten years of data, shows that ozone levels are higher in the summer months and that concentrations during those months have decreased with time.

Perhaps unexpectedly, insight into bivariate data can benefit from an apparent separation of the two attributes. For example, each of two attributes can be assigned to a separate histogram, as shown in Figure 3.20(a). A single house is represented once on each histogram, as illustrated by the highlighted house in Figure 3.20(b). If the histograms are static there is no opportunity to see the relation between the two attributes and there is little to commend the use of two

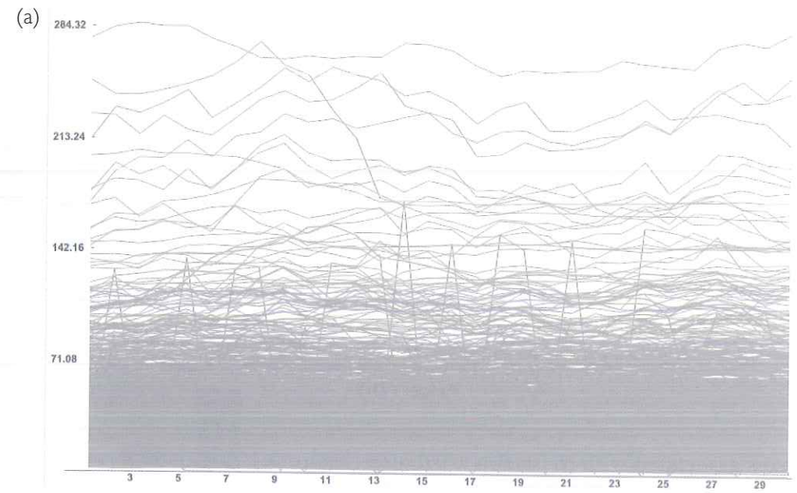


FIGURE 3.18 Four views of a time-series query tool. (a) An overview of the entire data set; (b) a single timebox limits the display to items with prices between $70 and $250 during days 1 to 4; (c) an additional constraint selects items with prices between $70 and $95 during days 7 to 12; (d) yet another constraint concerns prices between $90 and $115 for days 15 to 18 *(Courtesy of Harry Hochheiser)*

separate histograms. However, if interaction allows the placement of limits on one of the attribute scales (Figure 3.20(c)), then the houses thereby identified can usefully be encoded by colour not only on one attribute histogram but on the other as well (Figure 3.20(c)), thereby providing insight into the relation between the two attributes. This is another example of the powerful technique of brushing, whose value is difficult to overstate. Especially if the histogram display is modified to 'range down' selected objects, as in Figure 3.20(d), another benefit accruing from brushing is illustrated if it is possible to move the entire selected *range* of one attribute from side to side and see how that affects the selected houses on the second attribute. If lower and upper limits on both attribute scales can be adjusted separately, a very flexible exploratory visualization tool results (Tweedie *et al.,* 1994; Spence and Tweedie, 1998; Albinsson *et al.,* 2003).

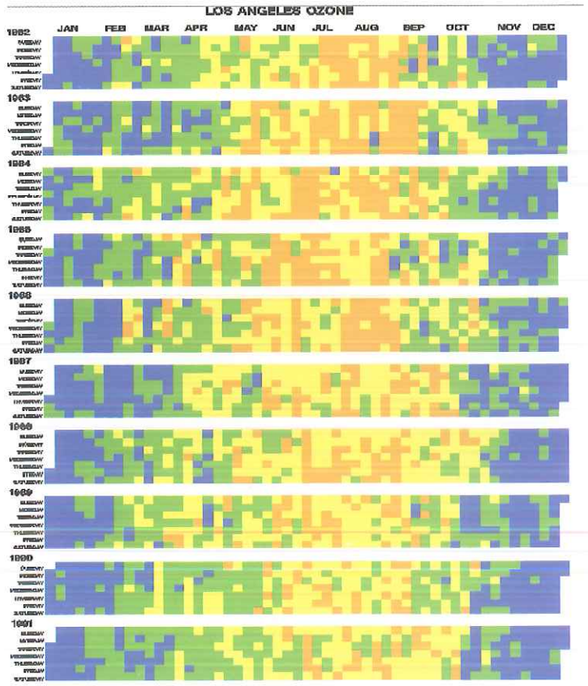


FIGURE 3.19
Representation of the level of ozone concentration above Los Angeles over a period of ten years

It may well be the case that, of two attributes, one is either far more important than the other or must be examined first. In this case it may be appropriate to employ 'logical' or *'semantic' zoom*. If the price of cars is of prime interest, a representation such as that shown in Figure 3.21(a) might be the first to be examined. It is then possible to arrange for a semantic zoom that will show, for a subset of the cars, the make of car in addition to the price (Figure 3.21(b)). This technique, which dates back to 1980 at least (Herot, 1980; Herot *et al.,* 1981), is quite general: it can encompass many attributes and many discrete levels of progressive zoom, as we shall see in the next chapter.

The frequent need for a *qualitative* understanding of data is illustrated in the representation of Figure 3.22. The domain is that of electronic circuit design and the situation is one in which the designer has proposed a design and needs a first appreciation of a particular property of that circuit, for example the magnitude of the voltage at each point in the circuit. These values are encoded by the size of a red square in Figure 3.22. Why? First because the designer will already have a mental model of the expected voltage magnitudes and the display will either confirm that model or, in the case of Figure 3.22, make it clear that a mistake has been made in choosing a value for a particular component (because the two squares at top right are not of equal area). In this example the designer has made a useful discovery, one that might not have been made if the voltage values had

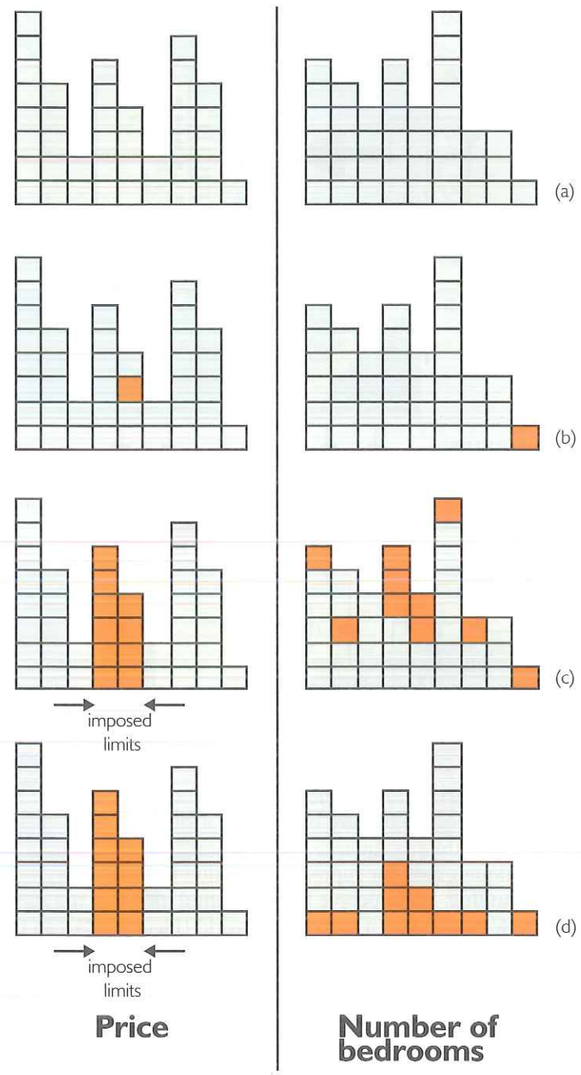


FIGURE 3.20
Linked histograms. (a) The price and number of bedrooms associated with a collection of houses are represented by separate histograms; (b) a single house is represented once on each histogram; (c) upper and lower limits placed on price define a subset of houses which are coded red on *both* histograms; (d) interpretation is enhanced by 'ranging down' the colour-coded houses, especially if exploration involves the dynamic alteration of limits