Preface

This book is about using interactive and dynamic plots on a computer screen as part of data exploration and modeling, both alone and as a partner with static graphics and non-graphical computational methods. The area of interactive and dynamic data visualization emerged within statistics as part of research on exploratory data analysis in the late 1960s, and it remains an active subject of research today, as its use in practice continues to grow. It now makes substantial contributions within computer science as well, as part of the growing fields of information visualization and data mining, especially visual data mining.

The material in this book includes:

- An introduction to data visualization, explaining how it differs from other types of visualization.
- A description of our toolbox of interactive and dynamic graphical methods.
- An approach for exploring missing values in data.
- An explanation of the use of these tools in cluster analysis and supervised classification.
- An overview of additional material available on the web.
- A description of the data used in the analyses and exercises.

The book's examples use the software R and GGobi. R (Ihaka & Gentleman 1996, R Development Core Team 2006) is a free software environment for statistical computing and graphics; it is most often used from the command line, provides a wide variety of statistical methods, and includes high–quality static graphics. R arose in the Statistics Department of the University of Auckland and is now developed and maintained by a global collaborative effort. It began as a re-implementation of the S language and statistical computing environment (Becker & Chambers 1984) first developed at Bell Laboratories before the breakup of AT&T.

GGobi (Swayne, Temple Lang, Buja & Cook 2003) is free software for interactive and dynamic graphics; it can be operated using a command-line interface or from a graphical user interface (GUI). When GGobi is used as a stand-alone tool, only the GUI is used; when it is used with R, via the `rggobi` (Temple Lang, Swayne, Wickham & Lawrence 2006) package, a command-line interface is used along with the GUI. GGobi is a descendant of two earlier programs: XGobi (Swayne, Cook & Buja 1992, Buja, Cook & Swayne 1996) and, before that, Dataviewer (Buja, Hurley & McDonald 1986, Hurley 1987). Many of the examples that follow might be reproduced with other software such as S-PLUS®, SAS JMP®, DataDesk®, Mondrian, MANET, and Spotfire®. However, GGobi is unique because it offers tours (rotations of data in higher than 3D), complex linking between plots using categorical variables, and the tight connection with R.

Web resources

The web site which accompanies the book contains sample datasets and R code, movies demonstrating the interactive and dynamic graphic methods, and additional chapters. It can be reached through the GGobi web site:

`http://www.ggobi.org`

The R software is available from:

`http://www.R-project.org`

Both web sites include source code as well as binaries for various operating systems (Linux®, Windows®, Mac OS X®) and allow users to sign up for mailing lists and browse mailing list archives. The R web site offers a wealth of documentation, including an introduction to R and a partially annotated list of books offering more instruction. [Widely read books include Dalgaard (2002), Venables & Ripley (2002), and Maindonald & Braun (2003).] The GGobi web site includes an introductory tutorial, a list of papers, and several movies.

How to use this book

The language in the book is aimed at later year undergraduates, beginning graduate students, and graduate students in any discipline needing to analyze their own multivariate data. It is suitable reading for an industry statistician, engineer, bioinformaticist, or computer scientist with some knowledge of basic data analysis and a need to analyze high-dimensional data. It also may be useful for a mathematician who wants to visualize high-dimensional structures.

The end of each chapter contains exercises to help practice the methods discussed in the chapter. The book may be used as a text in a class on statistical graphics, exploratory data analysis, visual data mining, or information visualization. It might also be used as an adjunct text in a course on multivariate data analysis or data mining.

This book has been closely tied to a particular software implementation so that you can actively use the methods as you read about them, to learn and experiment with interactive and dynamic graphics. The plots and written explanations in the book are no substitute for personal experience. We strongly urge the reader to go through this book sitting near a computer with

GGobi, R, and `rggobi` installed, following along with the examples. If you do not wish to install the software, then the next best choice is to watch the accompanying movies demonstrating the examples in the text.

If you have not used GGobi before, then visit the web site, watch the movies, download the manual, and work through the tutorial; the same advice applies for those unfamiliar with R: Visit the R web site and learn the basics.

As you read the book, try things out for yourself. Take your time, and have fun!

Acknowledgments

This work started at Bellcore (now Telcordia), where Jon Kettering, Ram Gnanadesikan, and Diane Duffy carefully nurtured graphics research in the 1990s. It has continued with the encouragement of our respective employers at AT&T (with support from Daryl Pregibon and Chris Volinsky), at Lucent Bell Laboratories (with support from Diane Lambert), and at Iowa State University (with support from chairs Dean Isaacson and Ken Koehler). Werner Stuetzle provided the initial encouragement to write this book. A large part of the book was drafted while Dianne Cook was partially supported for a semester by Macquarie University, Sydney.

The authors would like to thank the peer reviewers recruited by the publisher for their many useful comments.

Proceeds from the sale of each book will be directed to the GGobi Foundation, to nurture graphics research and development.

Dianne Cook
Iowa State University

Deborah F. Swayne
AT&T Labs – Research

July 2007

Contents

Technical Notes

R code

The R code in this book, denoted by `typewriter font`, and the more extensive code on the web site, has been tested on version 2.4.0 of R, version 2.1.5 of GGobi, and version 2.1.5 of `rggobi`. Updates will be available on the web site as they are needed.

Figures

The figures in this book were produced in a variety of ways, and the files and code to reproduce them are all available on the book's web site. Some were produced directly in R. Some were produced using both GGobi and R, and the process of converting GGobi views into publication graphics deserves an explanation.

When we arrive at a GGobi view we want to include in a paper or book, we use the `Save Display Description` item on GGobi's `Tools` menu to generate a file containing an S language description of the display. We read the file into R using the R package `DescribeDisplay` (Wickham 2006*b*), like this:

```
> library(DescribeDisplay)
> d <- dd_load("fig.R")
```

We create the publication-quality graphic using either that package's plot method or another R package, `ggplot` (Wickham 2006*c*), like this:

```
> plot(d)
```

or

```
> p <- ggplot(d)
> print(p)
```

Figure 0.1 illustrates the differences with a trio of representations of the same bivariate scatterplot. The picture at left is a screen dump of a GGobi display. Such images are not usually satisfactory for publication for several

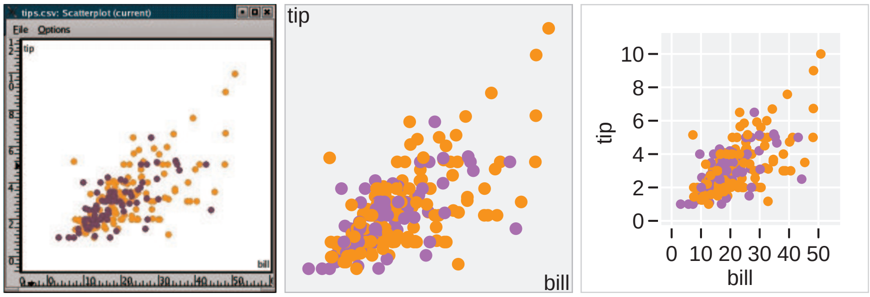

Fig. 0.1. Sample plots produced from GGobi in different ways: **(left)** a simple screen dump; **(middle)** a plot produced using the plot method of the R package `DescribeDisplay`; **(right)** a plot made using the R package `ggplot`.

reasons, the most obvious of which is the lack of resolution. The second picture was produced using `DescribeDisplay`'s plot method, which reproduces the plotting region of the view with pretty good fidelity. We used this method to produce most of the one–dimensional and two-dimensional tour pictures in this book. The third picture was produced using `ggplot`, which adds axis ticks, labels and grid lines. We used it to produce nearly all the bivariate scatterplots of GGobi views in this book.

List of Figures

1

Introduction

In this technological age, we live in a sea of information. We face the problem of gleaning useful knowledge from masses of words and numbers stored in computers. Fortunately, the computing technology that produces this deluge also gives us some tools to transform heterogeneous information into knowledge. We now rely on computers at every stage of this transformation: structuring and exploring information, developing models, and communicating knowledge.

In this book we teach a methodology that makes visualization central to the process of abstracting knowledge from information. Computers give us great power to represent information in pictures, but even more, they give us the power to interact with these pictures. If these are pictures of data, then interaction gives us the feeling of having our hands on the data itself and helps us to orient ourselves in the sea of information. By generating and manipulating many pictures, we make comparisons among different views of the data, we pose queries about the data and get immediate answers, and we discover large patterns and small features of interest. These are essential facets of data exploration, and they are important for model development and diagnosis.

In this first chapter we sketch the history of computer-aided data visualization and the role of data visualization in the process of data analysis.

1.1 Data visualization: beyond the third dimension

So far we have used the terms "information," "knowledge," and "data" informally. From now on we will use the following distinction: the term *data* refers to information that is structured in some schematic form such as a table or a list, and knowledge is derived from studying data. Data is often but not always quantitative, and it is often derived by processing unstructured information. It always includes some attributes or variables such as the number of hits on web sites, frequencies of words in text samples, weight in pounds, mileage in gallons per mile, income per household in dollars, years of education, acidity on the pH scale, sulfur emissions in tons per year, or scores on standardized tests.

When we visualize data, we are interested in portraying abstract relationships among such variables: for example, the degree to which income increases with education, or whether certain astronomical measurements indicate grouping and therefore hint at new classes of celestial objects. In contrast to this interest in abstract relationships, many other areas of visualization are principally concerned with the display of objects and phenomena in physical three-dimensional (3D) space. Examples are volume visualization (e.g., for the display of human organs in medicine), surface visualization (e.g., for manufacturing cars or animated movies), flow visualization (e.g., for aeronautics or meteorology), and cartography (e.g., for navigation or social studies). In these areas one often strives for physical realism or the display of great detail in space, as in the visual display of a new car design or of a developing hurricane in a meteorological simulation. The data visualization task is obviously different from drawing physical objects.

If data visualization emphasizes abstract variables and their relationships, then the challenge of data visualization is to create pictures that reflect these abstract entities. One approach to drawing abstract variables is to create axes in space and map the variable values to locations on the axes, and then render the axes on a drawing surface. In effect, one codes non-spatial information using spatial attributes: position and distance on a page or computer screen. The goal of data visualization is then not realistic drawing, which is meaningless in this context, but translating abstract relationships to interpretable pictures.

This way of thinking about data visualization, as interpretable spatial representation of abstract data, immediately brings up a limitation: Plotting surfaces such as paper or computer screens are merely two-dimensional (2D). We can extend this limit by simulating a third dimension: The eye can be tricked into seeing 3D virtual space with perspective and motion, but if we want an axis for each variable, that's as far as we can stretch the display dimension.

This limitation to a 3D display space is not a problem if the objects to be represented are three-dimensional, as in most other visualization areas. In data visualization, however, the number of axes required to code variables can be large: Five or ten axes are common, but these days one often encounters dozens and even hundreds. Overcoming the 2D and 3D barriers is a key challenge for data visualization. To meet this challenge, we use powerful computer-aided visualization tools. For example, we can mimic and amplify a strategy familiar from photography: taking pictures from multiple directions so the shape of an object can be understood in its entirety. This is an example of the "multiple views" paradigm, which will be a recurring theme of this book. In our 3D world the paradigm works superbly, because the human eye is adept at inferring the true shape of an object from just a few directional views. Unfortunately, the same is often not true for views of abstract data. The chasm between different views of data, however, can be actively bridged with additional computer technology: Unlike the passive paper medium, computers allow us to manipulate pictures, to pull and push their content in continuous motion like a moving video camera, or to poke at objects in one picture and see them light up in other pictures. Motion links pictures in time; poking links them across space. This book features many illustrations of the power of these linking technologies. The diligent reader may come away "seeing" high-dimensional data spaces!

1.2 Statistical data visualization: goals and history

Visualization has been used for centuries to map our world (cartography) and describe the animal and plant kingdoms (scientific illustration). Data visualization, which is more abstract, emerged more recently. An early innovator was William Playfair, whose extensive charting of economic data in the 1800s (Wainer & Spence 2005*a*, Wainer & Spence 2005*b*) contributed to its emergence. The early history of visualization has been richly — and beautifully — documented (Friendly & Denis 2006, Tufte 1983, Tufte 1990, Ford 1992, Wainer 2000).

Today's data visualization has homes in several disciplines, including the natural sciences, engineering, geography, computer science, and statistics. There is a lot of overlap in the functionality of the methods and tools they generate, but some differences in emphasis can be traced to the research contexts in which they were incubated. For example, data visualization in the natural science and engineering communities supports the goal of modeling physical objects and processes, relying on scientific visualization (Hansen & Johnson 2004, Bonneau, Ertl & Nielson 2006). For the geographical community, maps are the starting point, and other data visualization methods are used to expand on the information displayed using cartography (Longley, Maguire, Goodchild & Rhind 2005, Dykes, MacEachren & Kraak 2005). The database research community creates visualization software that grows out of their work in data storage and retrieval; their graphics often summarize the kinds of tables and tabulations that are common results of database queries. The human–computer interface community produces software as part of their research in human perception, human–computer interaction and usability, and their tools are often designed to make the performance of a complex task as straightforward as possible. These two latter fields have been instrumental in developing the field of information visualization (Card, Mackinlay & Schneiderman 1999, Bederson & Schneiderman 2003, Spence 2007).

The statistics community creates visualization systems within the context of data analysis, so the graphics are designed to support and enrich the statistical processes of data exploration, modeling, and inference. As a result, statistical data visualization has some unique features. Statisticians are always concerned with variability in observations and error in measurements, both of which cause uncertainty about conclusions drawn from data. Dealing with this uncertainty is at the heart of classical statistics, and statisticians have developed a huge body of inferential methods that help to quantify uncertainty.

Systems for data analysis included visualization as soon as they began to emerge (Nie, Jenkins, Steinbrenner & Bent 1975, Becker & Chambers 1984, Wilkinson 1984). They could generate a wide variety of plots, either for display on the screen or for printing, and the more flexible systems have always allowed users considerably leeway in plot design. Since these systems predated the general use of the mouse, the keyboard was their only input device, and the displays on the screen were not themselves interactive.

As early as the 1960s, however, researchers in many disciplines were making innovations in computer–human interaction, and statisticians were there, too. The seminal visualization system for exploratory data analysis was PRIM-9, the work of Fisherkeller, Friedman, and Tukey at the Stanford Linear Accelerator Center in 1974. PRIM-9 was the first stab at an interactive tool set for the visual analysis of multivariate data. It was followed by further pioneering systems at the Swiss Federal Institute of Technology (PRIM-ETH), Harvard University (PRIM-H), and Stanford University (ORION), in the late 1970s and early 1980s.

Research picked up in the following few years in many places (Wang 1978, McNeil 1977, Velleman & Velleman 1985, Cleveland & McGill 1988, Buja & Tukey 1991, Tierney 1991, Cleveland 1993, Rao 1993, Carr, Wegman & Luo 1996). The authors themselves were influenced by work at Bell Laboratories, Bellcore, the University of Washington, Rutgers University, the University of Minnesota, MIT, CMU, Batelle Richmond WA, George Mason University, Rice University, York University, Cornell University, Trinity College, and the University of Augsburg, among others. In the past couple of years, books have begun to appear that capture this history and continue to point the way forward (Wilkinson 2005, Young, Valero-Mora & Friendly 2006, Unwin, Theus & Hofmann 2006, Chen, Härdle & Unwin 2007).

1.3 Getting down to data

Here is a very small and seemingly simple dataset we will use to illustrate the use of data graphics. One waiter recorded information about each tip he received over a period of a few months working in one restaurant. He collected several variables:

- tip (i.e., gratuity) in US dollars
- bill (the cost of the meal) in US dollars
- sex of the bill payer
- whether the party included smokers
- day of the week
- time of day
- size of the party

In all he recorded 244 tips. The data was reported in a collection of case studies for business statistics (Bryant & Smith 1995). The primary question suggested by the data is this: *What are the factors that affect tipping behavior?*

This dataset is typical (albeit small): There are seven variables, of which two are numeric (tip, bill), and the others are categorical or otherwise discrete. In answering the question, we are interested in exploring relationships that may involve more than three variables, none of which corresponds to physical space. In this sense the data is high-dimensional and abstract.

We look first at the variable of greatest interest to the waiter: tip. A common graph for looking at a single variable is the histogram, where data values are binned and the count is represented by a rectangular bar. We choose an initial bin width of $1 and produce the uppermost graph of Fig. 1.1. The distribution appears to be unimodal; that is, it has one peak, the bar representing the tips greater than $1.50 and less than or equal $2.50. There are very few tips of $1.50 or less. The number of larger tips trails off rapidly, which suggests that this is not a very expensive restaurant.

The conclusions drawn from a histogram are often influenced by the choice of bin width, which is a parameter of the graph and not of the data. Figure 1.1 shows a histogram with a smaller bin width, 10c. At the smaller bin width the shape is multimodal, and it is clear that there are large peaks at the full dollars and smaller peaks at the half dollar. This shows that the customers tended to round the tip to the nearest fifty cents or dollar.

This type of observation occurs frequently when studying histograms: A large bin width smooths out the graph and shows rough or global trends, whereas a smaller bin width highlights more local features. Since the bin width is an example of a graph parameter, experimenting with bin width is an example of exploring a set of related graphs. Exploring multiple related graphs can lead to insights that would not be apparent in any single graph.

So far we have not addressed the primary question: What relationships exist between tip and the other variables? Since the tip is usually calculated based on the bill, it is natural to look first at a graph of tip and bill. A common graph for looking at a pair of continuous variables is the scatterplot, as in Fig. 1.2. We see that the variables are highly correlated ($r = 0.68$), which confirms that tip is calculated from the bill. We have added a line representing a tip rate of 18%. Disappointingly for the waiter, there are many more points below the line than above it: There are many more "cheap tippers" than generous tippers. There are a couple of notable exceptions, especially one

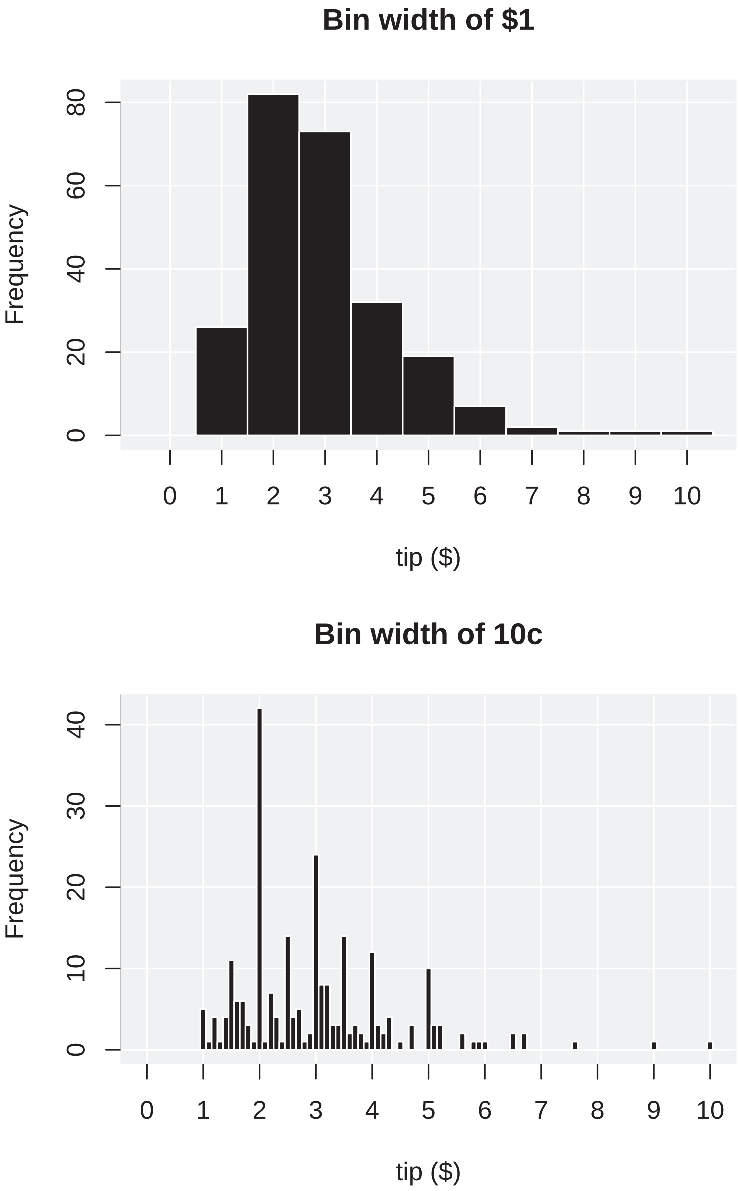

Fig. 1.1. Histograms of `tip` with differing bin width: $1, 10*c*. Bin width can be changed interactively in interactive systems, often by dragging a slider.

party who gave a $5.15 tip for a $7.25 bill, which works out to a tip rate of about 70%.

We said earlier that an essential aspect of data visualization is capturing relationships among many variables: three, four, or even more. This dataset, simple as it is, illustrates the point. Let us ask, for example, how a third variable such as sex affects the relationship between tip and bill. As sex is categorical with two levels (i.e., binary), it is natural to divide the data into female and male payers and to generate two scatterplots of tip vs. bill. Let us go even further by including a fourth variable, smoking, which is also binary. We now divide the data into four parts and generate the four scatterplots observed in Fig. 1.3. (The 18% tip guideline is included in each plot, and the correlation between the variables for each subset is in the top left of each plot.) Inspecting these plots reveals numerous features: (1) For smoking parties, there is a lot less association between the size of the tip and the size of the bill; (2) when a female non-smoker paid the bill, the tip was a very consistent percentage of the bill, with the exceptions of three dining parties; and (3) larger bills were mostly paid by men.

Taking stock

In the above example we gained a wealth of insight in a short time. Using nothing but graphical methods we investigated univariate, bivariate, and multivariate relationships. We found both global features and local detail. We saw that tips were rounded; then we saw the obvious correlation between the tip and the size of the bill, noting the scarcity of generous tippers; finally we discovered differences in the tipping behavior of male and female smokers and non-smokers.

Notice that we used very simple plots to explore some pretty complex relationships involving as many as four variables. We began to explore multivariate relationships for the first time when we produced the plots in Fig. 1.3. Each plot shows a subset obtained by partitioning the data according to two binary variables. The statistical term for partitioning based on variables is "conditioning." For example, the top left plot shows the dining parties that meet the condition that the bill payer was a male non-smoker: sex = male and smoking = False. In database terminology this plot would be called the result of "drill-down." The idea of conditioning is richer than drill-down because it involves a structured partitioning of *all* data as opposed to the extraction of a single partition.

Having generated the four plots, we arrange them in a two-by-two layout to reflect the two variables on which we conditioned. Although the axes in each plot are tip and bill, the axes of the overall figure are smoking (vertical) and sex (horizontal). The arrangement permits us to make several kinds of comparisons and to make observations about the partitions. For example, comparing the rows shows that smokers and non-smokers differ in the strength of the correlation between tip and bill, and comparing the plots in the top row shows that male and female non-smokers differ in that the larger bills tend

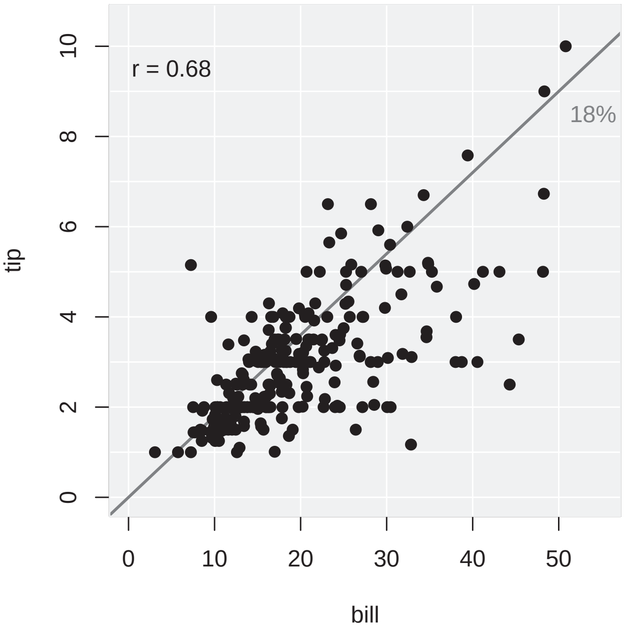

Fig. 1.2. Scatterplot of tip vs. bill. The line represents a tip of 18%. The greater number of points far below the line indicates that there are more "cheap tippers" than generous tippers.

to be paid by men. In this way a few simple plots allow us to reason about relationships among four variables.

In contrast, an old-fashioned approach without graphics would be to fit a regression model. Without subtle regression diagnostics (which rely on graphics!), this approach would miss many of the above insights: the rounding of tips, the preponderance of cheap tippers, and perhaps the multivariate relationships involving the bill payer's sex and the group's smoking habits.

1.4 Getting real: process and caveats

The preceding explanations may have given a somewhat misleading impression of the process of data analysis. In our account the data had no problems; for example, there were no missing values and no recording errors. Every step was logical and necessary. Every question we asked had a meaningful answer. Every plot that was produced was useful and informative. In actual data

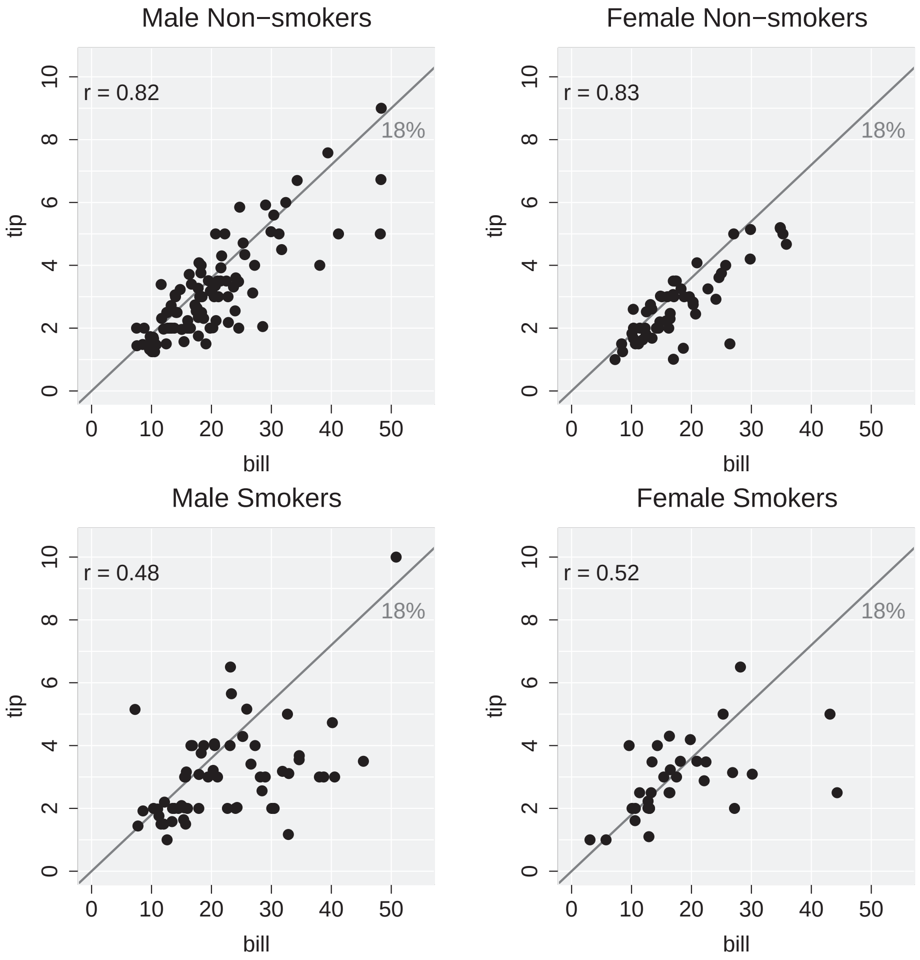

Fig. 1.3. Scatterplot of tip vs. bill conditioned by sex and smoker. There is almost no association between tip and bill in the smoking parties, and with the exception of three dining parties, when a female non-smoker paid the bill, the tip was extremely consistent.

analysis, nothing could be further from the truth. Real datasets are rarely perfect; most choices are guided by intuition, knowledge, and judgment; most steps lead to dead ends; most plots end up in the wastebasket. This may sound daunting, but even though data analysis is a highly improvisational activity, it can be given some structure nonetheless.

To understand data analysis, and how visualization fits in, it is useful to talk about it as a process consisting of several stages:

- The problem statement
- Data preparation
- Exploratory data analysis
- Quantitative analysis
- Presentation

The problem statement: Why do you want to analyze this data? Underlying every dataset is a question or problem statement. For the tipping data the question was provided to us from the data source: "What are the factors that affect tipping behavior?" This problem statement drives the process of any data analysis. Sometimes the problem is identified prior to a data collection. Perhaps it is realized after data becomes available because having the data available has made it possible to imagine new issues. It may be a task that the boss assigns, it may be an individual's curiosity, or it may be part of a larger scientific endeavor. Ideally, we begin an analysis with some sense of direction, as described by a pertinent question.

Data preparation: In the classroom, the teacher hands the class a single data matrix with each variable clearly defined. In the real world, it can take a great deal of work to construct a clean data matrix. For example, data values may be missing or misrecorded, data may be distributed across several sources, and the variable definitions and data values may be inconsistent across these sources. Analysts often have to invest considerable time in acquiring domain knowledge and in learning computing tools before they can even ask a meaningful question about the data. It is therefore not uncommon for this stage to consume most of the effort that goes into a project. And it is also not uncommon to loop back to this stage after completing the subsequent stages, to re-prepare and re-analyze the data.

In preparing the Tips data, we would create a new variable called tip rate, because when tips are discussed in restaurants, among waiters, dining parties, and tourist guides, it is in terms of a percentage of total bill. We may also create several new dummy variables for the day of the week, in anticipation of fitting a regression model. We did not talk about using visualization to verify that we had correctly understood and prepared the tipping data. For example, that unusually large tip could have been the result of a transcription error. Graphics identified the observation as unusual, and the analyst might use this information to search the origins of the data to check the validity of the numbers for this observation.

Exploratory data analysis (EDA): At this stage in the analysis, we make time to "play in the sand" to allow us to find the unexpected, and come to some understanding of our data. We like to think of this as a little like traveling. We may have a purpose in visiting a new city, perhaps to attend a conference, but we need to take care of our basic necessities, such as finding eating places and gas stations. Some of our movements will be pre-determined, or guided by the advice of others, but some of the time we wander around by ourselves.

We may find a cafe we particularly like or a cheaper gas station. This is all about getting to know the neighborhood.

By analogy, at this stage in the analysis, we relax the focus on the problem statement and explore broadly different aspects of the data. Modern exploratory data analysis software is designed to make this process as fruitful as possible. It is a highly interactive, real-time, dynamic, and visual process, having evolved along with computers. It takes advantage of technology, in a way that Tukey envisioned and experimented with on specialist hardware 40 years ago: "Today, software and hardware together provide far more powerful factories than most statisticians realize, factories that many of today's most able young people find exciting and worth learning about on their own" (Tukey 1965). It is characterized by direct manipulation and dynamic graphics: plots that respond in real time to an analyst's queries and change dynamically to re-focus, link to information from other sources, and re-organize information. The analyst can work rapidly and thoroughly through the data, slipping out of dead-ends and chasing down new leads. The high level of interactivity is enabled by bare-bones graphics, which are generally not adequate for presentation purposes.

We gave you some flavor of this stage in the analysis of the waiter's tips. Although the primary question was about the factors affecting tipping behavior, we checked the distribution of individual variables, we looked for unusual records, we explored relationships among multiple variables, and we found some unexpected patterns: the rounding of tips, the prevalence of cheap tippers, and the heterogeneity in variance between groups.

Quantitative analysis (QA):

At this stage, we use statistical modeling and statistical interference to answer our primary questions. With statistical models, we summarize complex data, decomposing it into estimates of signal and noise. With statistical inference, we try to assess whether a signal is real. Data visualization plays an important role at this stage, although that is less well known than its key role in exploration. It is helpful both in better understanding a model and in assessing its validity in relation to the data.

For Tips, we have not yet answered the primary question of interest. Let's fit a regression model using tiprate as the response and the remaining variables (except tip and bill) as the explanatory variables. When we do this, only size has a significant regression coefficient, resulting in the model $\widehat{tiprate} = 0.18 - 0.01 \times size$. The model says that, starting from a baseline tip rate of 18%, the amount drops by 1% for each additional diner in a party, and this is the model answer in Bryant & Smith (1995). Figure 1.4 shows this model and the underlying data. (The data is *jittered* horizontally to alleviate overplotting caused by the discreteness of size; that is, a small amount of noise is added to the value of size for each case.)

Are we satisfied with this model? We have some doubts about it, although we know that something like it is used in practice: Most restaurants today factor the tip into the bill automatically for larger dining parties. However, in this data it explains only 2% of the variation in tip rate. The points are spread widely around the regression line. There are very few data points for parties of size one, five, and six, which makes us question the validity of the model in these regions. The signal is very weak relative to the noise.

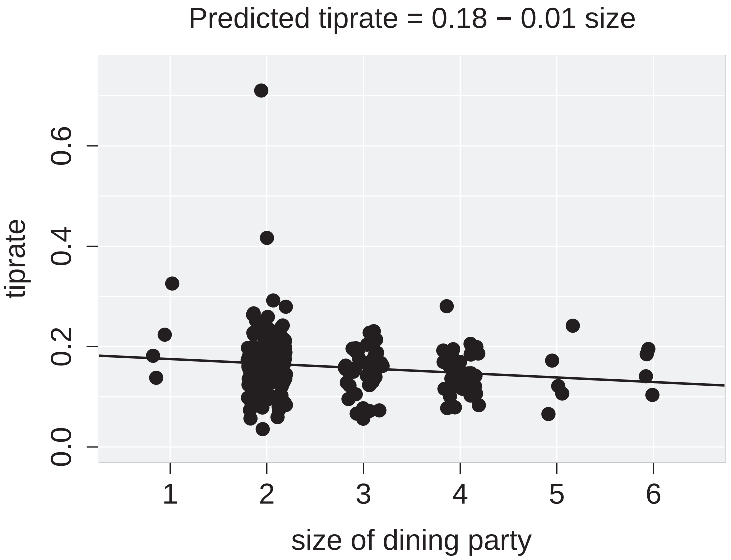

Fig. 1.4. Factors affecting tipping behavior. This scatterplot of tiprate vs. size shows the best model along with the data (jittered horizontally). There is a lot of variation around the regression line, showing very little signal relative to noise. In addition there are very few data points for parties of 1, 5, or 6 diners, so the model may not be valid at these extremes.

Most problems are more complex than the Tips data, and the models are often more sophisticated, so evaluating them is correspondingly more difficult. We evaluate a model using data produced by the model-fitting process, such as model estimates and diagnostics. Other data may be derived by simulating from the model or by calculating confidence regions. All this data can be explored and plotted for the pleasure of understanding the model.

Plotting the model in relation to the original data is also important. There is a temptation to ignore that messy raw data in favor of the simplification provided by a model, but a lot can be learned from what is left out of a model. For example, we would never consider teaching regression analysis without teaching residual plots. A model is a succinct explanation of the variation in the data, a simplification. With a model we can make short descriptive statements about the data, and pictures help us find out whether a model is *too* simple. And so we plot the model in the context of the data, as we just did in Fig. 1.4, and as we will do often in the chapters to follow.

The interplay of EDA and QA: Is it data snooping?

Because EDA is very graphical, it sometimes gives rise to a suspicion that patterns in the data are being detected and reported that are not really there. Sometimes this is called *data snooping*. Certainly it is important to validate our observations about the data. Just as we argue that models should be validated by all means available, we are just as happy to argue that observations made in plots should be validated using quantitative methods, permutation tests, or cross-validation, as appropriate, and incorporating subject matter expertise. A discussion of this topic emerged in the comments on Koschat & Swayne (1996), and Buja's remark (Buja 1996) is particularly apt:

> In our experience, false discovery is the lesser danger when compared to nondiscovery. Nondiscovery is the failure to identify meaningful structure, and it may result in false or incomplete modeling. In a healthy scientific enterprise, the fear of nondiscovery should be at least as great as the fear of false discovery.

We snooped into the Tips data, and from a few plots we learned an enormous amount of information about tipping: There is a scarcity of generous tippers, the variability in tips increases extraordinarily for smoking parties, and people tend to round their tips. These are very different types of tipping behaviors than what we learned from the regression model. The regression model was not compromised by what we learned from graphics, and indeed, we have a richer and more informative analysis. Making plots of the data is just smart.

On different sides of the pond: EDA and IDA

Consulting statisticians, particularly in the British tradition, have always looked at the data before formal modeling, and call it IDA (initial data analysis) (Chatfield 1995). For example, Crowder & Hand (1990) say: “The first thing to do with data is to look at them.... usually means tabulating and plotting the data in many different ways to ‘see what's going on’. With the wide availability of computer packages and graphics nowadays there is no excuse for ducking the labour of this preliminary phase, and it may save some red faces later.”

The interactive graphics methods described in this book emerged from a different research tradition, which started with Tukey's influential work on EDA, focusing on discovery and finding the unexpected in data. Like IDA, EDA has always depended heavily on graphics, even before the term *data visualization* was coined. Our favorite quote from John Tukey's rich legacy is that we need good pictures to “force the unexpected upon us.”

EDA and IDA, although not entirely distinct, differ in emphasis. Fundamental to EDA is the desire to let the data inform us, to approach the data without pre-conceived hypotheses, so that we may discover unexpected features. Of course, some of the unexpected features may be errors in the data. IDA emphasizes finding these errors by checking the quality of data prior to formal modeling. It is much more closely tied to inference than EDA: Problems with the data that violate the assumptions required for valid inference need to be discovered and fixed early.

In the past, EDA and inference were sometimes seen as incompatible, but we argue that they are not mutually exclusive. In this book, we present some visual methods for assessing uncertainty and performing inference, that is, deciding whether what we see is "really there."

Presentation: Once an analysis has been completed, the results must be reported, either to clients, managers, or colleagues. The results probably take the form of a narrative and include quantitative summaries such as tables, forecasts, models, and graphics. Quite often, graphics form the bulk of the summaries.

The graphics included in a final report may be a small fraction of the graphics generated for exploration and diagnostics. Indeed, they may be different graphics altogether. They are undoubtedly carefully prepared for their audience. The graphics generated during the analysis are meant for the analyst only and thus need to be quickly generated, functional but not polished. This issue is a dilemma for authors who have much to say about exploratory graphics but need to convey it in printed form. The plots in this book, for example, lie somewhere between exploratory and presentation graphics.

As mentioned, these broadly defined stages do not form a rigid recipe. Some stages overlap, and occasionally some are skipped. The order is often shuffled and groups of steps reiterated. What may look like a chaotic activity is often improvisation on a theme loosely following the "recipe."

Because of its improvisational nature, EDA is not easy to teach. Says Tukey (1965) "Exploratory data analysis is NOT a bundle of techniques....Confirmatory analysis is easier to teach and compute...." In the classroom, the teacher explains a method to the class and demonstrates it on the single data matrix and then repeats this process with another method. Teaching a bundle of methods is indeed an efficient approach to covering substantial quantities of material, but this may be perceived by the student as a stream of disconnected methods, applied to unrelated data fragments, and they may not be able to apply what they have learned outside that fragmented context for quite a while. It takes time and experience for students to integrate this material and to develop their own intuition. Students need to navigate their own way through data, cleaning it, exploring it, choosing models; they need to make mistakes, recover from them, and synthesize the findings into a summary. Learning how to perform data analysis is a process that continues long after the student's formal training is complete.

1.5 Interactive investigation

Thus far, all observations on the tipping data have been made using static graphics — our purpose up to this point has been to communicate the importance of plots in the context of data analysis. Static plots were originally drawn by hand, and although they are now produced by computers, they are still designed to be printed on paper, often to be displayed or studied some time later. However, computers also allow us to produce plots to be viewed as they are created, and tweaked and manipulated in real time. This book is about such interactive and dynamic plots, and the chapters that follow have a lot to say about them. Here we will say a few words about the way interactive plots enhance the data analysis process we have just described.

The Tips data is simple, and most of the interesting features can be discovered using static plots. Still, interacting with the plots reveals more and enables the analyst to pursue follow-up questions. For example, we could address a new question, arising from the current analysis, such as "Is the rounding behavior of tips predominant in some demographic group?" To investigate we probe the histogram, highlight the bars corresponding to rounded tips, and observe the pattern of highlighting in the linked plots (Fig. 1.5). Multiple plots are visible simultaneously, and the highlighting action on one plot generates changes in the other plots. The two additional plots here are *mosaic plots*, which are used to examine the proportions in categorical variables. (Mosaic plots will be explained further in the next chapter; for now, it is enough to know that the area of each rectangle is proportional to the corresponding number of cases in the data.) For the highlighted subset of dining parties, the ones who rounded the tip to the nearest dollar or half-dollar, the proportion of bill paying males and females is roughly equal, but interestingly, the proportion of smoking parties is higher than non-smoking parties. This might suggest another behavioral difference between smokers and non-smokers: a larger tendency for smokers than non-smokers to round their tips. If we were to be skeptical about this effect we would dig deeper, making more graphical explorations and numerical models. By pursuing this with graphics, we would find that the proportion of smokers who round the tip is only higher than non-smokers for full dollar amounts, and not for half-dollar amounts.

The remaining chapters in this book continue in this vein, describing how interactive and dynamic plots are used in several kinds of data analysis.

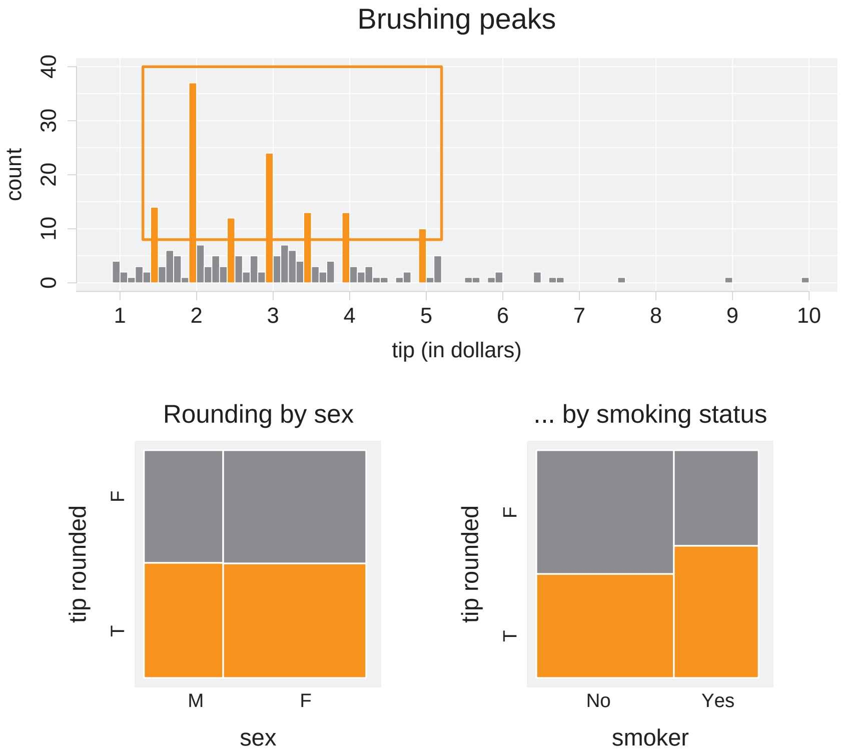

Fig. 1.5. Histogram of tip linked to 2D mosaic plots of sex and smoker. Bars of whole and half-dollar amounts are highlighted. The proportion of smoking parties who round their tips is higher than that of non-smoking parties, whereas men and women round about equally.