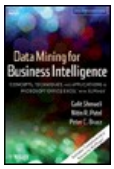

Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner, Second Edition

by Galit Shmueli, Nitin R. Patel and Peter C. Bruce

Chapter 3: Data Visualization

In this chapter we describe a set of plots that can be used to explore the multidimensional nature of a dataset. We present basic plots (bar charts, line graphs, and scatterplots), distribution plots (boxplots and histograms), and different enhancements that expand the capabilities of these plots to visualize more information. We focus on how the different visualizations and operations can support data mining tasks, from supervised (prediction, classification, and time series forecasting) to unsupervised tasks, and provide some guidelines on specific visualizations to use with each data mining task. We also describe the advantages of interactive visualization over static plots. The chapter concludes with a presentation of specialized plots that are suitable for data with special structure (hierarchical, network, and geographical).

3.1 Uses of Data Visualization

The popular saying "a picture is worth a thousand words" refers to the ability to condense diffused verbal information into a compact and quickly understood graphical image. In the case of numbers, data visualization and numerical summarization provide us with both a powerful tool to explore data and an effective way to present results.

Where do visualization techniques fit into the data mining process, as described so far? Their use is primarily in the preprocessing portion of the data mining process. Visualization supports data cleaning by finding incorrect values (e.g., patients whose age is 999 or —1), missing values, duplicate rows, columns with all the same value, and the like. Visualization techniques are also useful for variable derivation and selection: they can help determine which variables to include in the analysis and which might be redundant. They can also help with determining appropriate bin sizes, should binning of numerical variables be needed (e.g., a numerical outcome variable might need to be converted to a binary variable, as was done in the Boston housing data, if a yes/no decision is required). They can also play a role in combining categories as part of the data reduction process. Finally, if the data have yet to be collected and collection is expensive (as with the Pandora project at its outset, see Chapter 7), visualization methods can help determine, using a sample, which variables and metrics are useful.

In this chapter we focus on the use of graphical presentations for the purpose of *data exploration*, in particular with relation to predictive analytics. Although our focus is not on visualization for the purpose of data reporting, this chapter offers ideas as to the effectiveness of various graphical displays for the purpose of data presentation. These offer a wealth of useful presentations beyond tabular summaries and basic bar charts, currently the most popular form of data presentation in the business environment. For an excellent discussion of using graphs to report business data, see Few (2004). In terms of reporting data mining results graphically, we describe common graphical displays elsewhere in the book, some of which are technique specific [e.g., dendrograms for hierarchical clustering (Chapter 14), and tree charts for classification and regression trees (Chapter 9)] while others are more general [e.g., receiver operating characteristic (ROC) curves and lift charts for classification (Chapter 5) and profile plots for clustering (Chapter 14)].

Data exploration is a mandatory initial step whether or not more formal analysis follows. Graphical exploration can support free-form exploration for the purpose of understanding the data structure, cleaning the data (e.g., identifying unexpected gaps or "illegal" values), identifying outliers, discovering initial patterns (e.g., correlations among variables and surprising clusters), and generating interesting questions. Graphical exploration can also be more focused, geared toward specific questions of interest. In the data mining context a combination is needed: free-form exploration performed with the purpose of supporting a specific goal.

Graphical exploration can range from generating very basic plots to using operations such as filtering and zooming interactively to explore a set of interconnected plots that include advanced features such as color and multiple panels. This chapter is not meant to be an exhaustive guidebook on visualization techniques but instead discusses main principles and features that support data exploration in a data mining context. We start by describing varying levels of sophistication in terms of visualization and show the advantages of different features and operations. Our discussion is from the perspective of how visualization supports the subsequent data mining goal. In particular, we distinguish between supervised and unsupervised learning; within supervised learning, we also further distinguish between classification (categorical *Y*) and prediction (numerical *Y*).

3.2 Data Examples

To illustrate data visualization we use two datasets that are used in additional chapters in the book. This allows the reader to compare some of the basic Excel plots used in other chapters to the improved plots and easily see the merit of advanced visualization.

Example 1: Boston Housing Data

The Boston housing data contain information on census tracts in Boston for which several measurements are taken (e.g., crime rate, pupil/teacher ratio). It has 14 variables (a description of each variable and the data are given in Chapter 2, in Table 2.2 and Figure 2.5). We consider three possible tasks:

1. A supervised predictive task, where the outcome variable of interest is the median value of a home in the tract (MEDV)
2. A supervised classification task, where the outcome variable of interest is the binary variable CAT.MEDV that equals 1 for tracts with median home value above $30, 000 and equals 0 otherwise
3. An unsupervised task, where the goal is to cluster census tracts

(MEDV and CAT.MEDV are not used together in any of the three cases.)

Example 2: Ridership on Amtrak Trains

Amtrak, a U.S. railway company, routinely collects data on ridership. Here we focus on forecasting future ridership using the series of monthly ridership between January 1991 and March 2004. The data and their source are described in Chapter 15. Hence our task here is (numerical) time series forecasting.

3.3 Basic Charts: Bar Charts, Line Graphs, and Scatterplots

The three most effective basic plots are bar charts, line graphs, and scatterplots. These plots are easy to create in Microsoft Excel and are the most commonly used in the current business world, in both data exploration and presentation (unfortunately, pie charts are also popular, although usually ineffective visualizations). Basic charts support data exploration by displaying one or two columns of data (variables) at a time. This is useful in the early stages of getting familiar with the data structure, the amount and types of variables, the volume and type of missing values, and the like.

The nature of the data mining task and domain knowledge about the data will affect the use of basic charts in terms of the amount of time and effort allocated to different variables. In supervised learning, there will be more focus on the outcome variable. In scatterplots, the outcome variable is typically associated with the *y* axis. In unsupervised learning (for the purpose of data reduction or clustering), basic plots that convey relationships (such as scatterplots) are preferred.

The top left panel in Figure 3.1 displays a line chart for the time series of monthly railway passengers on Amtrak. Line graphs are used primarily for showing time series. The choice of time frame to plot, as well as the temporal scale, should depend on the horizon of the forecasting task and on the nature of the data.

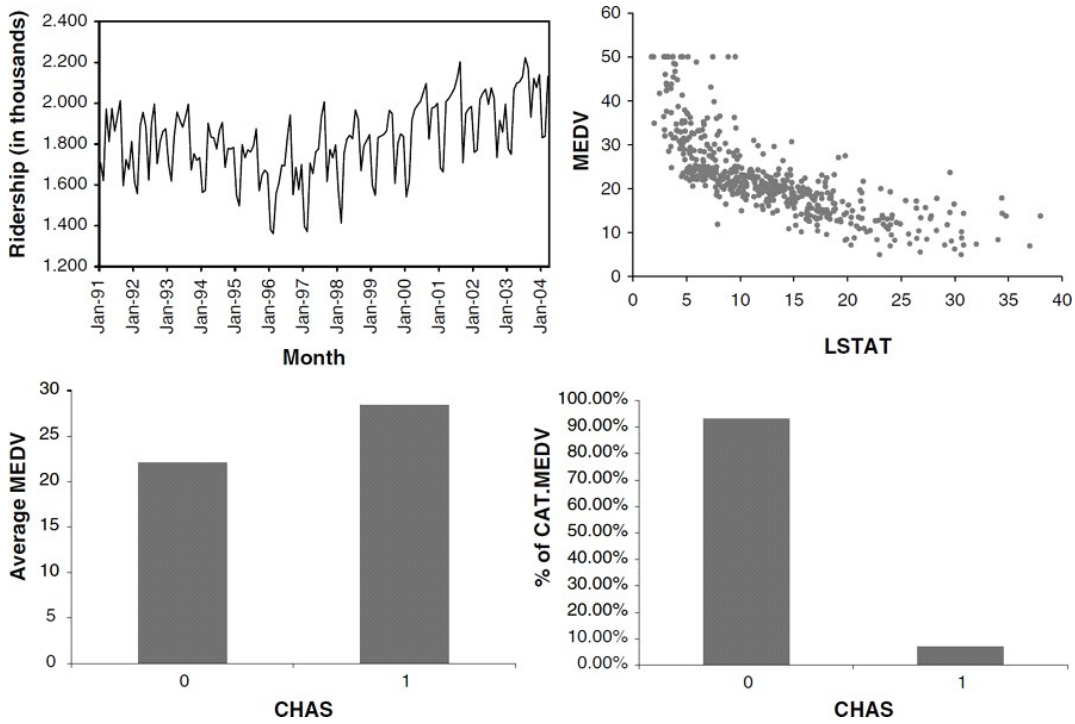

FIGURE 3.1: BASIC PLOTS: LINE GRAPH (TOP LEFT), SCATTERPLOT (TOP RIGHT), BAR CHART FOR NUMERICAL VARIABLE (BOTTOM LEFT), AND BAR CHART FOR CATEGORICAL VARIABLE (BOTTOM RIGHT)

Bar charts are useful for comparing a single statistic (e.g., average, count, percentage) across groups. The height of the bar (or length, in a horizontal display) represents the value of the statistic, and different bars correspond to different groups. Two examples are shown in the bottom panels in Figure 3.1. The left panel shows a bar chart for a numerical variable (MEDV) and the right panel shows a bar chart for a categorical variable (CAT.MEDV). In each, separate bars are used to denote homes in Boston that are near the Charles River versus those that are not (thereby comparing the two categories of CHAS). The chart with the numerical output MEDV (bottom left) uses the average MEDV on the *y* axis. This supports the predictive task: The numerical outcome is on the *y* axis and the *x* axis is used for a potential categorical predictor.[1] (Note that the *x* axis on a bar chart must be used only for categorical variables because the order of bars in a bar chart should be interchangeable.) For the classification task, CAT.MEDV is on the *y* axis (bottom right), but its aggregation is a percentage (the alternative would be a count). This graph shows us that the vast majority (over 90%) of the tracts do not border the Charles River (CHAS=0). Note that the labeling of the *y* axis can be confusing in this case: the value of CAT.MEDV plays no role and the *y* axis is simply a percentage of all records.

The top right panel in Figure 3.1 displays a scatterplot of MEDV versus LSTAT. This is an important plot in the prediction task. Note that the output MEDV is again on the *y* axis (and LSTAT on the *x* axis is a potential predictor). Because both variables in a basic scatterplot must be numerical, it cannot be used to display the relation between CAT.MEDV and potential predictors for the classification task (but we can enhance it to do so—see Section 3.4). For unsu-pervised learning, this particular scatterplot helps study the association between two numerical variables in terms of information overlap as well as identifying clusters of observations.

All three basic plots highlight global information such as the overall level of ridership or MEDV, as well as changes over time (line chart), differences between subgroups (bar chart), and relationships between numerical variables (scatterplot).

Distribution Plots: Boxplots and Histograms

Before moving on to more sophisticated visualizations that enable multidimensional investigation, we note two important plots that are usually not considered "basic charts" but are very useful in statistical and data mining contexts. The *boxplot* and the *histogram* are two plots that display the entire distribution of a numerical variable. Although averages are very popular and useful summary statistics, there is usually much to be gained by looking at additional statistics such as the median and standard deviation of a variable, and even more so by examining the entire distribution. Whereas bar charts can only use a single aggregation, boxplots and histograms display the entire distribution of a numerical variable. Boxplots are also effective for comparing subgroups by generating side-by-side boxplots, or for looking at distributions over time by creating a series of boxplots.

Distribution plots are useful in supervised learning for determining potential data mining methods and variable transformations. For example, skewed numerical variables might warrant transformation (e.g., moving to a logarithmic scale) if used in methods that assume normality (e.g., linear regression, discriminant analysis).

A histogram represents the frequencies of all *x* values with a series of vertical connected bars. For example, in the top left panel of Figure 3.2, there are about 20 tracts where the median value (MEDV) is between $7500 and $12, 500.

A boxplot represents the variable being plotted on the *y* axis (although the plot can potentially be turned in a 90° angle, so that the boxes are parallel to the *x* axis). In the top right panel of Figure 3.2 there are two boxplots (called a side-by-side boxplot). The box encloses 50% of the data—for example, in the right-hand box half of the tracts have median values (MEDV) between $20, 000 and $33, 000. The horizontal line inside the box represents the median (50th percentile). The top and bottom of the box represent the 75th and 25th percentiles, respectively. Lines extending above and below the box cover the rest of the data range; outliers may be depicted as points or circles. Sometimes the average is marked by a + (or similar) sign, as in the top right panel of Figure 3.2. Comparing the average and the median helps in assessing how skewed the data are. Boxplots are often arranged in a series with a different plot for each of the various values of a second variable, shown on the *x* axis.

Because histograms and boxplots are geared toward numerical variables, in their basic form they are useful for prediction tasks. Boxplots can also support unsupervised learning by displaying relationships between a numerical variable (*y* axis) and a categorical variable (*x* axis). To illustrate these points, see Figure 3.2. The top panel shows a histogram of MEDV, revealing a skewed distribution. Transforming the output variable to log(MEDV) would likely improve results of a linear regression predictor.

The right panel in Figure 3.2 shows side-by-side boxplots comparing the distribution of MEDV for homes that border the Charles River (1) or not (0), (similar to Figure 3.1). We see that not only is the average MEDV for river-bounding homes higher than the non-river-bounding homes, the entire distribution is higher (median, quartiles, min, and max). We also see that all river-bounding homes have MEDV above $10, 000, unlike non-river-bounding homes. This information is useful for identifying the potential importance of this predictor (CHAS) and for choosing data mining methods that can capture the nonoverlapping area between the two distributions (e.g., trees). Boxplots and histograms applied to numerical variables can also provide directions for deriving new variables, for example, they can indicate how to bin a numerical variable (e.g., binning a numerical outcome in order to use a naive Bayes classifier, or in the Boston housing example, choosing the cutoff to convert MEDV to CAT.MEDV).

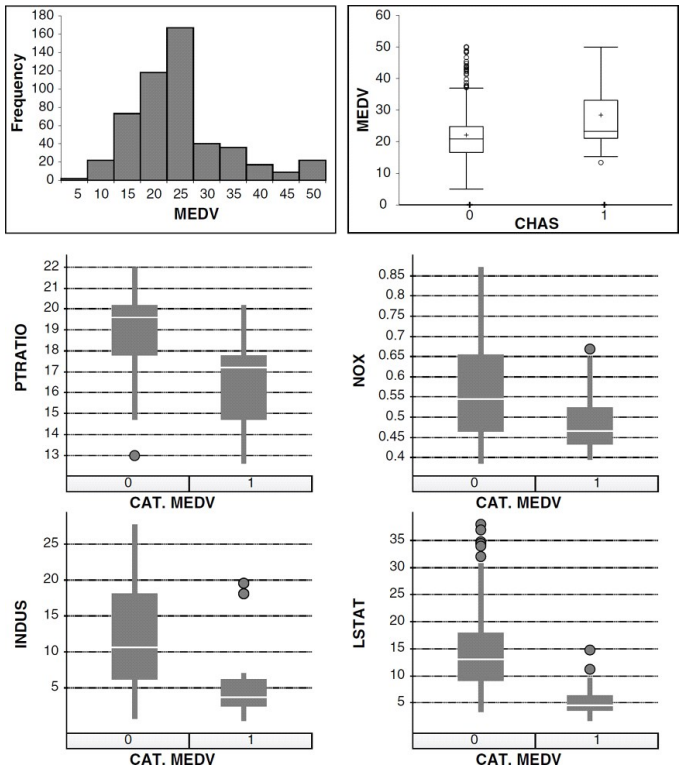

FIGURE 3.2: EXAMPLES OF HISTOGRAM (TOP LEFT) AND SIDE-BY-SIDE BOXPLOTS CREATED WITH XLMINER (TOP RIGHT) AND SPOTFIRE (CENTER AND BOTTOM ROWS). NOTE THAT IN A SIDE-BY-SIDE BOXPLOT, ONE AXIS IS USED FOR A CATEGORICAL VARIABLE, AND THE OTHER FOR A NUMERICAL VARIABLE. A NUMERICAL OUTCOME VARIABLE, IF IT IS PLOTTED, WILL APPEAR ON THE CATEGORICAL AXIS (IN WHICH CASE WE ARE PLOTTING THE DISTRIBUTION OF ONE OF THE NUMERICAL PREDICTORS). A NUMERICAL OUTCOME VARIABLE, IF IT IS PLOTTED, WILL APPEAR ON THE NUMERICAL AXIS (IN WHICH CASE WE ARE PLOTTING THE DISTRIBUTION OF THE OUTCOME VARIABLE ITSELF, WITH A CATEGORICAL PREDICTOR ON THE CATEGORICAL AXIS)

Finally, side-by-side boxplots are useful in classification tasks for evaluating the potential of numerical predictors. This is done by using the *x* axis for the categorical outcome and the *y* axis for a numerical predictor. An example is shown in the center and bottom rows of Figure 3.2, where we can see the effects of four numerical predictors on CAT.MEDV. The pairs that are most separated (e.g., PTRATIO and INDUS) indicate potentially useful predictors.

Boxplots and histograms are not readily available in Microsoft Excel (although they can be constructed through a tedious manual process). They are available in a wide range of statistical software packages. In XLMiner they can be generated through the *Charts* menu (we note the current limitation of five categories for side-by-side boxplots).

The main weakness of basic charts and distribution plots, in their basic form (i.e., using position in relation to the axes to encode values), is that they can only display two variables and therefore cannot reveal high-dimensional information. Each of the basic charts has two dimensions, where each dimension is dedicated to a single variable. In data mining, the data are usually multivariate by nature, and the analytics are designed to capture and measure multivariate information. Visual exploration should therefore also incorporate this important aspect. In the next section we describe how to extend basic charts (and distribution charts) to multidimensional data visualization by adding features, employing manipulations, and

incorporating interactivity. We then present several specialized charts that are geared toward displaying special data structures (Section 3.5).

Heatmaps: Visualizing Correlations and Missing Values

A *heatmap* is a graphical display of numerical data where color is used to denote values. In a data mining context, heatmaps are especially useful for two purposes: for visualizing correlation tables and for visualizing missing values in the data. In both cases the information is conveyed in a two-dimensional table. A correlation table for *p* variables has *p* rows and *p* columns. A data table contains *p* columns (variables) and *n* rows (records). If the number of rows is huge, then a subset can be used. In both cases it is much easier and faster to scan the color coding rather than the values. Note that heatmaps are useful when examining a large number of values, but they are not a replacement for more precise graphical display, such as bar charts, because color differences cannot be perceived accurately.

An example of a correlation table heatmap is shown in Figure 3.3, showing all the pairwise correlations between 14 variables (MEDV and 13 predictors). Darker shades correspond to stronger (positive or negative) correlation. It is easy to quickly spot the high and low correlations. This heatmap was produced using Excel's *Conditional Formatting*.

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
CRIM	1.00													
ZN	-0.20	1.00												
INDUS	0.41	-0.53	1.00											
CHAS	-0.06	-0.04	0.06	1.00										
NOX	0.42	-0.52	0.76	0.09	1.00									
RM	-0.22	0.31	-0.39	0.09	-0.30	1.00								
AGE	0.35	-0.57	0.64	0.09	0.73	-0.24	1.00							
DIS	-0.38	0.66	-0.71	-0.10	-0.77	0.21	-0.75	1.00						
RAD	0.63	-0.31	0.60	-0.01	0.61	-0.21	0.46	-0.49	1.00					
TAX	0.58	-0.31	0.72	-0.04	0.67	-0.29	0.51	-0.53	0.91	1.00				
PTRATIO	0.29	-0.39	0.38	-0.12	0.19	-0.36	0.26	-0.23	0.46	0.46	1.00			
B	-0.39	0.18	-0.36	0.05	-0.38	0.13	-0.27	0.29	-0.44	-0.44	-0.18	1.00		
LSTAT	0.46	-0.41	0.60	-0.05	0.59	-0.61	0.60	-0.50	0.49	0.54	0.37	-0.37	1.00	
MEDV	-0.39	0.36	-0.48	0.18	-0.43	0.70	-0.38	0.25	-0.38	-0.47	-0.51	0.33	-0.74	1.00

FIGURE 3.3: HEATMAP OF A CORRELATION TABLE. DARKER VALUES DENOTE STRONGER CORRELATION

In a missing value heatmap rows correspond to records and columns to variables. We use a binary coding of the original dataset where 1 denotes a missing value and 0 otherwise. This new binary table is then colored such that only missing value cells (with value 1) are colored. Figure 3.4 shows an example of a missing value heatmap for a dataset with over 1000 columns. The data include economic, social, political and "well-being" information on different countries around the world (each row is a country). The variables were merged from multiple sources, and for each source information was not always available on every country. The missing data heatmap helps visualize the level and amount of "missingness" in the merged data file. Some patterns of "missingness" easily emerge: variables that are missing for nearly all observations, as well as clusters of rows (countries) that are missing many values. Variables with little missingness are also visible. This information can then be used for determining how to handle the missingness (e.g., dropping some variables, dropping some records, imputing, or via other techniques).

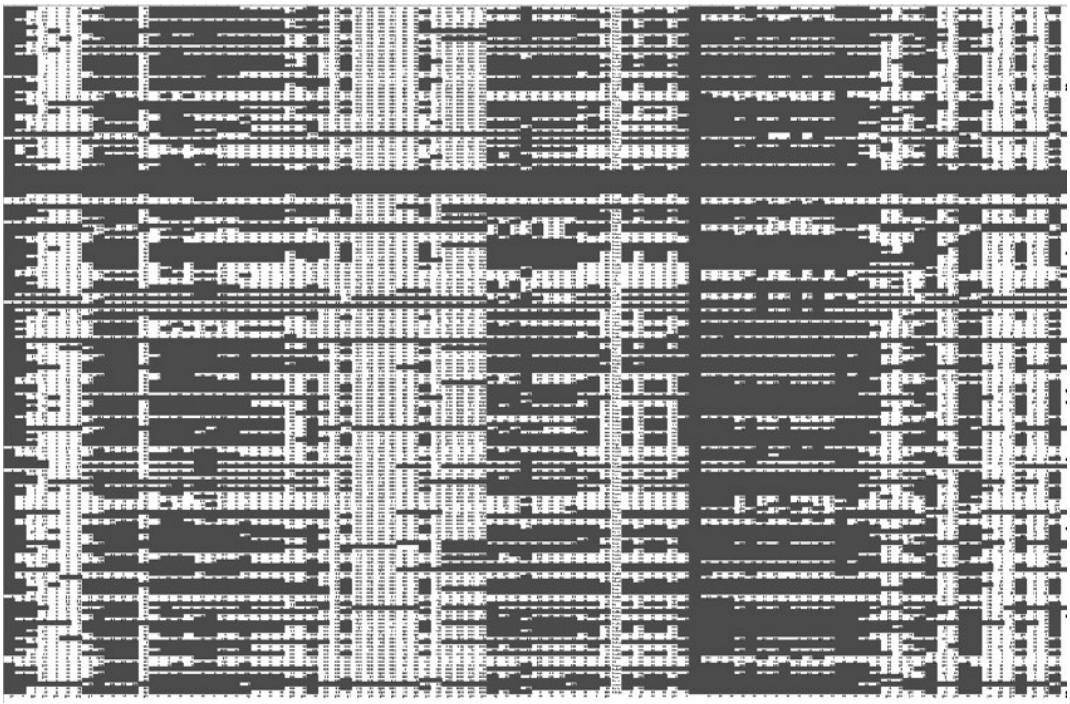

FIGURE 3.4: HEATMAP OF MISSING VALUES IN A DATASET. BLACK DENOTES MISSING VALUE

[1]We refer here to a bar chart with vertical bars. The same principles apply if using a bar chart with horizontal bars, except that the *x* axis is now associated with the numerical variable and the *y* axis with the categorical variable.

3.4 Multidimensional Visualization

Basic plots can convey richer information with features such as color, size, and multiple panels, and by enabling operations such as rescaling, aggregation, and interactivity. These additions allow looking at more than one or two variables at a time. The beauty of these additions is their effectiveness in displaying complex information in an easily understandable way. Effective features are based on understanding how visual perception works [see Few (2009) for a discussion]. The purpose is to make the information more understandable, not just represent the data in higher dimensions (such as three-dimensional plots that are usually ineffective visualizations).

Adding Variables: Color, Size, Shape, Multiple Panels, and Animation

In order to include more variables in a plot, we must consider the type of variable to include. To represent additional categorical information, the best way is to use hue, shape, or multiple panels. For additional numerical information we can use color intensity or size. Temporal information can be added via animation.

Incorporating additional categorical and/or numerical variables into the basic (and distribution) plots means that we can now use all of them for both prediction and classification tasks! For example, we mentioned earlier that a basic scatterplot cannot be used for studying the relationship between a categorical outcome and predictors (in the context of classification). However, a very effective plot for classification is a scatterplot of two numerical predictors color coded by the categorical outcome variable. An example is shown in the left panel of Figure 3.5, with color denoting CAT.MEDV.

In the context of prediction, color coding supports the exploration of the conditional relationship between the numerical outcome (on the *y* axis) and a numerical predictor. Color-coded scatterplots then help assess the need for creating interaction terms (e.g., is the relationship between MEDV and LSTAT different for homes near versus away from the river?).

Color can also be used to include further categorical variables into a bar chart, as long as the number of categories is small. When the number of categories is large, a better alternative is to use multiple panels. Creating multiple panels (also called "trellising") is done by splitting the observations according to a categorical variable and creating a separate plot (of the same type) for each category. An example is shown in the right panel of Figure 3.5, where a bar chart of average MEDV by RAD is broken down into two panels by CHAS. We see that the average MEDV for different highway accessibility levels (RAD) behaves differently for homes near the river (lower panel) compared to homes away from the river (upper panel). This is especially salient for RAD=1. We also see that there are no near-river homes in RAD levels 2, 6, and 7.

Such information might lead us to create an interaction term between RAD and CHAS and to consider condensing some of the bins in RAD. All these explorations are useful for prediction and classification.

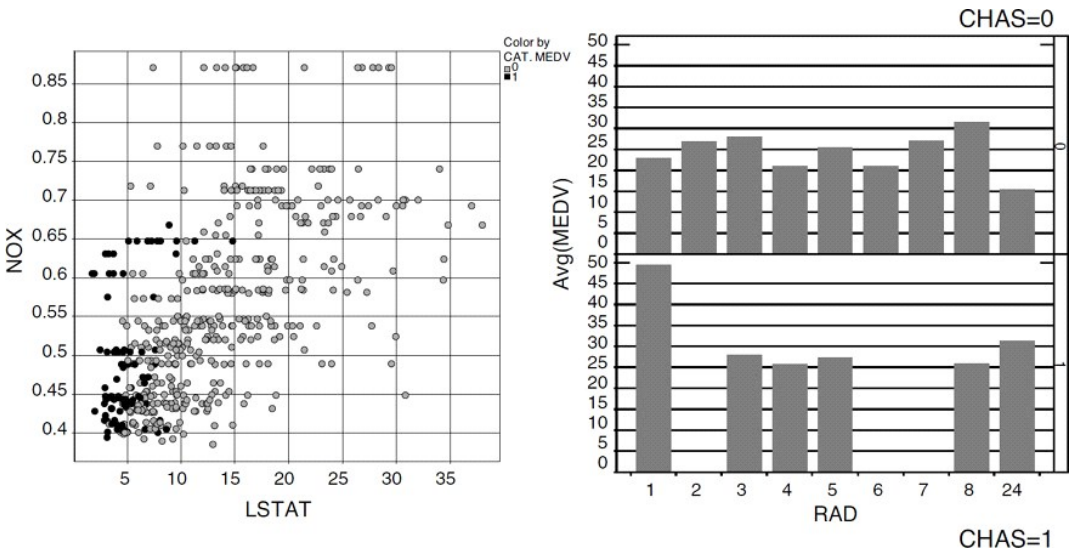

FIGURE 3.5: ADDING CATEGORICAL VARIABLES BY COLOR CODING AND MULTIPLE PANELS. (LEFT) SCATTERPLOT OF TWO NUMERICAL PREDICTORS, COLOR CODED BY THE CATEGORICAL OUTCOME (CAT.MEDV). (RIGHT) BAR CHART OF MEDV BY TWO CATEGORICAL PREDICTORS (CHAS AND RAD), USING MULTIPLE PANELS FOR CHAS. (CHAS = 0 FOR UPPER PANEL, CHAS = 1 FOR LOWER PANEL)

A special plot that uses scatterplots with multiple panels is the *scatterplot matrix*. In it, all pairwise scatterplots are shown in a single display. The panels in a scatterplot matrix are organized in a special way, such that each column corresponds to a variable and each row corresponds to a variable, thereby the intersections create all the possible pairwise scatterplots. The scatterplot matrix plot is useful in unsupervised learning for studying the associations between numerical variables, detecting outliers and identifying clusters. For supervised learning, it can be used for examining pairwise relationships (and their nature) between predictors to support variable transformations and variable selection (see Section 4.4). For prediction it can also be used to depict the relationship of the outcome with the numerical predictors.

An example of a scatterplot matrix is shown in Figure 3.6, with MEDV and three predictors. To identify which pair is plotted, variable names are shown along the diagonal cells; plots in the row corresponding to a variable show the variable's values along the *y* axis while plots in the corresponding column show the variable's values along the *x* axis. For example, the plots in the bottom row all have MEDV on the *y* axis (which allows studying the individual outcome-predictor relations). We can see different types of relationships from the different shapes (e.g., an exponential relationship between MEDV and LSTAT and a highly skewed relationship between CRIM and INDUS), which can indicate needed transformations. Note that the plots above and to the right of the diagonal are mirror images of those below and to the left.

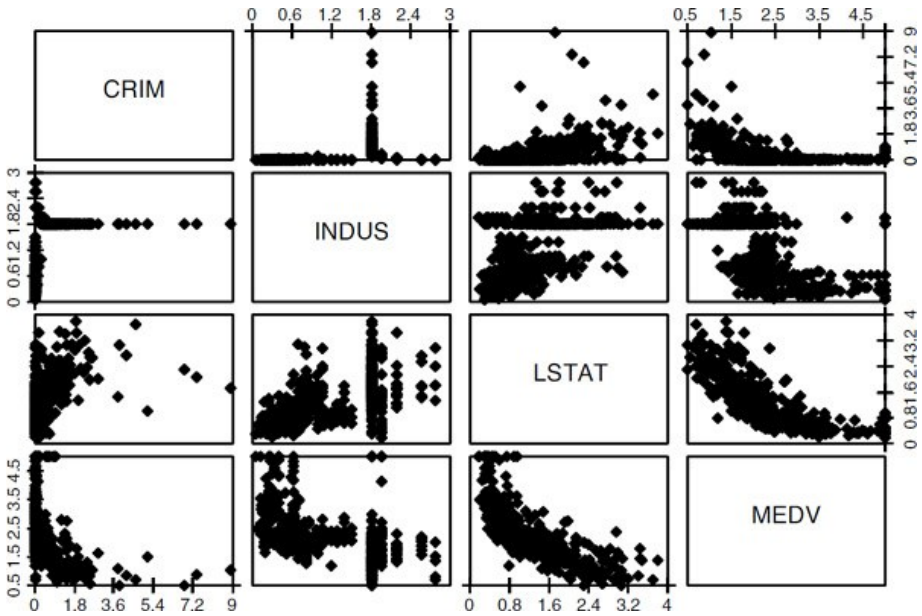

FIGURE 3.6: SCATTERPLOT MATRIX FOR MEDV AND THREE NUMERICAL PREDICTORS

Once hue is used, further categorical variables can be added via shape and multiple panels. However, one must proceed cautiously in adding multiple variables, as the display can become overcluttered and then visual perception is lost.

Adding a numerical variable via size is useful especially in scatterplots (thereby creating "bubble plots") because in a scatterplot, points represent individual observations. In plots that aggregate across observations (e.g., boxplots, histograms, bar charts) size and hue are not normally incorporated.

Finally, adding a temporal dimension to a plot to show how the information changes over time can be achieved via animation. A famous example is Rosling's animated scatterplots showing how world demographics changed over the years (www.gapminder.org). However, while animations of this type work for "statistical storytelling," they are not very effective for data exploration.

Manipulations: Rescaling, Aggregation and Hierarchies, Zooming, and Panning, and Filtering

Most of the time spent in data mining projects is spent in preprocessing. Typically, considerable effort is expended getting all the data in a format that can actually be used in the data mining software. Additional time is spent processing the data in ways that improve the performance of the data mining procedures. This preprocessing step in data mining includes variable transformation and derivation of new variables to help models perform more effectively. Transformations include changing the numeric scale of a variable, binning numerical variables, condensing categories in categorical variables, and the like. The following manipulations support the preprocessing step as well as the choice of adequate data mining methods. They do so by revealing patterns and their nature.

Rescaling Changing the scale in a display can enhance the plot and illuminate relationships. For example, in Figure 3.7 we see the effect of changing both axes of the scatterplot (top) and the *y* axis of a boxplot (bottom) to logarithmic (log) scale. Whereas the original plots (left) are hard to understand, the patterns become visible in log scale (right). In the scatterplot, the nature of the relationship between MEDV and CRIM is hard to determine in the original scale because too many of the points are "crowded" near the *y* axis. The rescaling removes this crowding and allows a better view of the linear relationship between the two log-scaled variables (indicating a log-log relationship). In the boxplot displays the crowding toward the *x* axis in the original units does not allow us to compare the two box sizes, their locations, lower outliers, and most of the distribution information. Rescaling removes the "crowding to the *x* axis" effect, thereby allowing a comparison of the two boxplots.

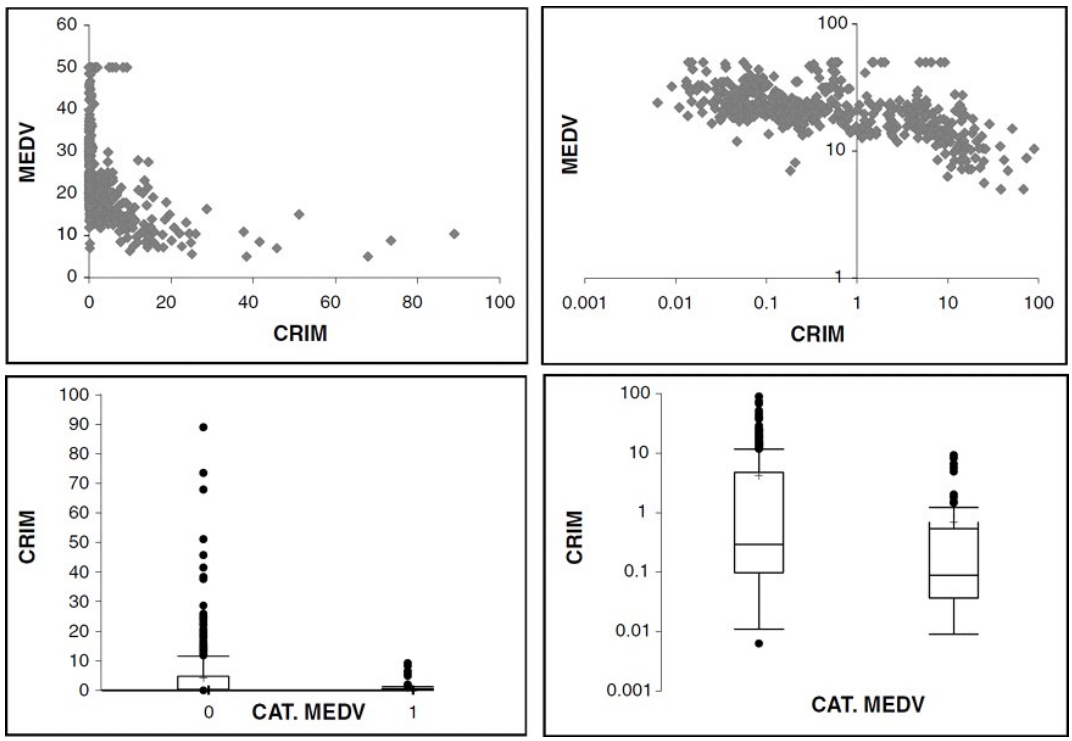

FIGURE 3.7: RESCALING CAN ENHANCE PLOTS AND REVEAL PATTERNS. (LEFT) ORIGINAL SCALE. (RIGHT) LOG SCALE

Aggregation and Hierarchies Another useful manipulation of scaling is changing the level of aggregation. For a temporal scale, we can aggregate by different granularity (e.g., monthly, daily, hourly) or even by a "seasonal" factor of interest such as month of year or day of week. A popular aggregation for time series is a moving average, where the average of neighboring values within a given window size is plotted. Moving-average plots enhance global trend visualization (see Chapter 15).

Nontemporal variables can be aggregated if some meaningful hierarchy exists: geographical (tracts within a zip code in the Boston housing example), organizational (people within departments within units), and so on. Figure 3.8 illustrates two types of aggregation for the railway ridership time series. The original monthly series is shown in the top left panel. Seasonal aggregation (by month of year) is shown in the top right panel, where it is easy to see the peak in ridership in July-Aug and the dip in Jan-Feb. The bottom right panel shows temporal aggregation, where the series is now displayed in yearly aggregates. This plot reveals the global long-term trend in ridership and the generally increasing trend from 1996 on.

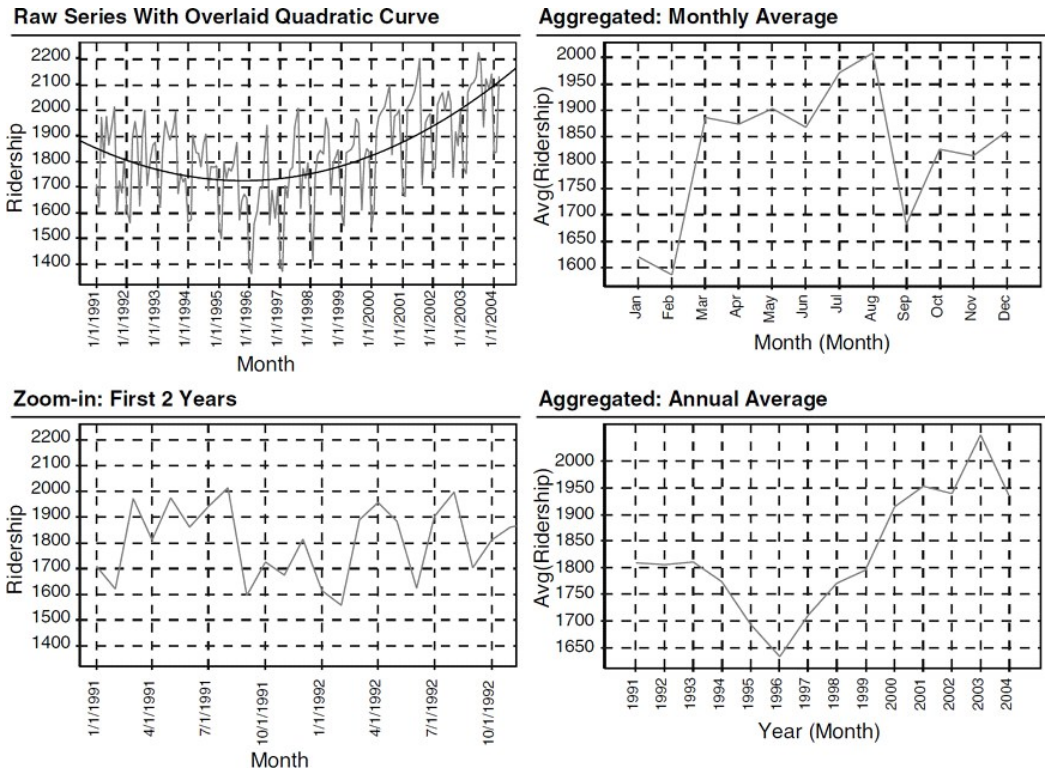

FIGURE 3.8: TIME SERIES LINE GRAPHS USING DIFFERENT AGGREGATIONS (RIGHT PANELS), ADDING CURVES (TOP LEFT PANEL), AND ZOOMING IN (BOTTOM LEFT PANEL). CREATED WITH SPOTFIRE

Examining different scales, aggregations, or hierarchies supports both supervised and unsupervised tasks in that it can reveal patterns and relationships at various levels and can suggest new sets of variables with which to work.

Zooming and Panning The ability to zoom in and out of certain areas of the data on a plot is important for revealing patterns and outliers. We are often interested in more detail on areas of dense information or of special interest. Panning refers to the operation of moving the zoom window to other areas (popular in mapping applications such as Google Maps). An example of zooming is shown in the bottom left panel of Figure 3.8, where the ridership series is zoomed in to the first 2 years of the series.

Zooming and panning support supervised and unsupervised methods by detecting areas of different behavior, which may lead to creating new interaction terms, new variables, or even separate models for data subsets. In addition, zooming and panning can help choose between methods that assume global behavior (e.g., regression models) and data-driven methods (e.g., exponential smoothing forecasters and *k*-nearest neighbors classifiers) and indicate the level of global-local behavior (as manifested by parameters such as *k* in *k*-nearest neighbors, the size of a tree, or the smoothing parameters in exponential smoothing).

Filtering Filtering means removing some of the observations from the plot. The purpose of filtering is to focus the attention on certain data while eliminating "noise" created by other data. Filtering supports supervised and unsupervised learning in a similar way to zooming and panning: It assists in identifying different or unusual local behavior.

Reference: Trend Lines and Labels

Trend lines and using in-plot labels also help to detect patterns and outliers. Trend lines serve as a reference and allow us to more easily assess the shape of a pattern. Although linearity is easy to visually perceive, more elaborate relationships such as exponential and polynomial trends are harder to assess by eye. Trend lines are useful in line graphs as well as in scatterplots. An example is shown in the top left panel of Figure 3.8, where a polynomial curve is overlaid on the original line graph (see also Chapter 15).

In displays that are not overcrowded, the use of in-plot labels can be useful for better exploration of outliers and clusters. An example is shown in Figure 3.9 (a reproduction of Figure 14.1 with the addition of labels). The figure shows different utilities on a scatterplot that compares fuel cost with total sales. We might be interested in clustering the data, and using clustering algorithms to identify clusters that differ markedly with respect to fuel cost and sales. Figure 14.1, with the labels,

helps visualize these clusters and their members (e.g., Nevada and Puget are part of a clear cluster with low fuel costs and high sales). For more on clustering, and this example, see Chapter 14.

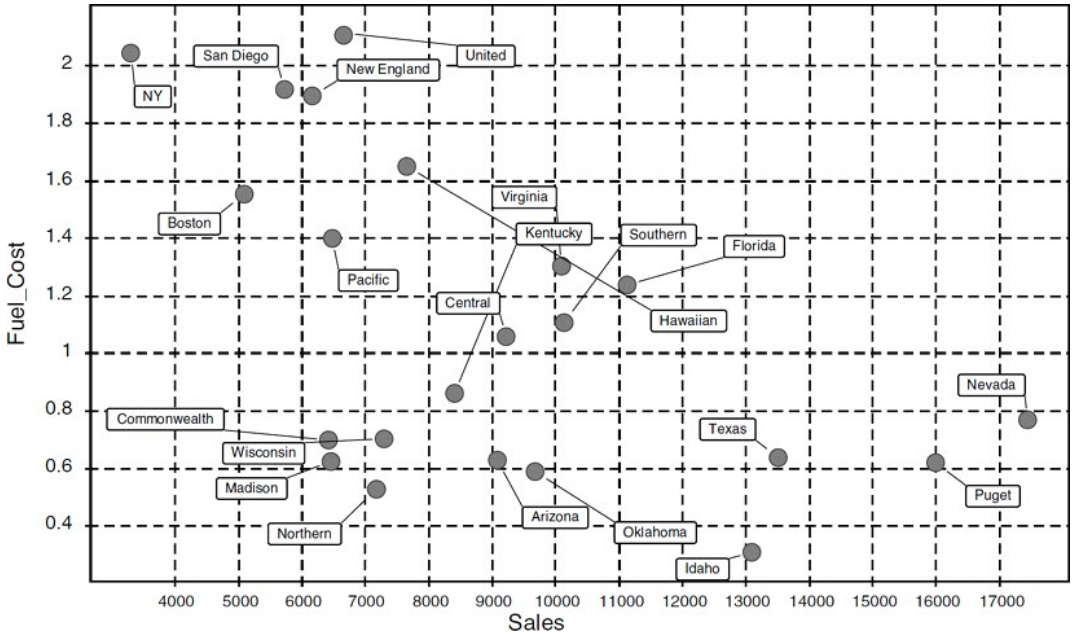

FIGURE 3.9: SCATTERPLOT WITH LABELED POINTS (CREATED WITH SPOTFIRE). COMPARE TO FIGURE 14.1

Scaling up: Large Datasets

When the number of observations (rows) is large, plots that display each individual observation (e.g., scatterplots) can become ineffective. Aside from applying aggregated charts such as boxplots, some alternatives are:

1. Sampling: drawing a random sample and using it for plotting (XLMiner has a sampling utility)
2. Reducing marker size
3. Using more transparent marker colors and removing fill
4. Breaking down the data into subsets (e.g., by creating multiple panels)
5. Using aggregation (e.g., bubble plots where size corresponds to number of observations in a certain range)
6. Using jittering (slightly moving each marker by adding a small amount of noise)

An example of the advantage of plotting a sample over the large dataset is shown in Figure 12.2 in Chapter 12, where a scatterplot of 5000 records is plotted alongside a scatterplot of a sample. Those plots were generated in Excel. We illustrate (Figure 3.10) an improved plot of the full dataset by applying smaller markers, using jittering to uncover overlaid points, and more transparent colors. We can see that larger areas of the plot are dominated by the gray class, the black class is mainly on the right, while there is a lot of overlap in the top right area.

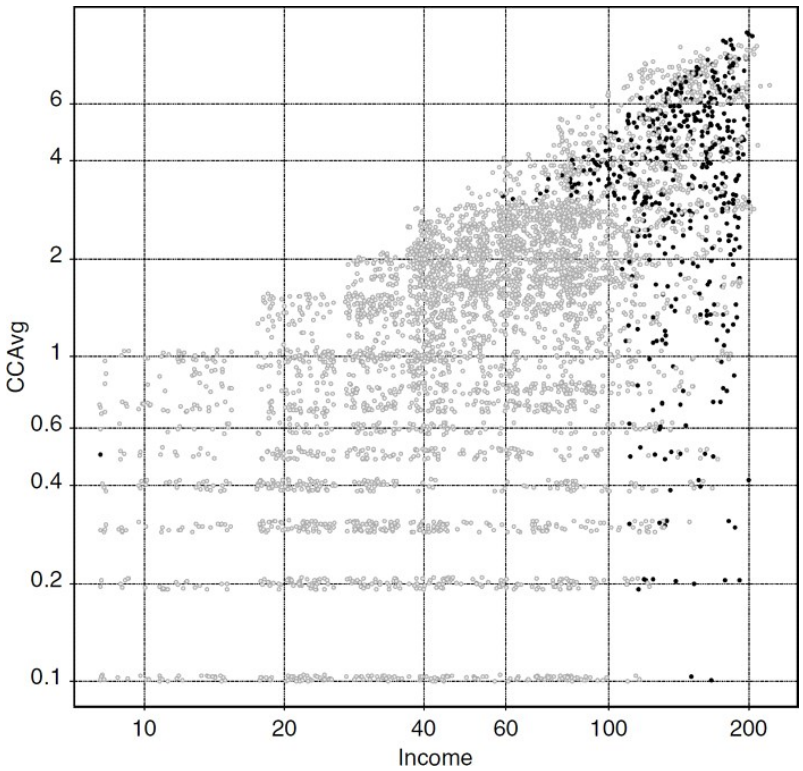

FIGURE 3.10: REPRODUCTION OF FIGURE I2.2 WITH REDUCED MARKER SIZE, JITTERING, AND MORE TRANSPARENT COLORING

Multivariate Plot: Parallel Coordinates Plot

Another approach toward presenting multidimensional information in a two-dimensional plot is via specialized plots such as the *parallel coordinates plot*. In this plot a vertical axis is drawn for each variable. Then each observation is represented by drawing a line that connects its values on the different axes, thereby creating a "multivariate profile." An example is shown in Figure 3.11 for the Boston housing data. In this display separate panels are used for the two values of CAT.MEDV in order to compare the profiles of homes in the two classes (for a classification task). We see that the more expensive homes (bottom panel) consistently have low CRIM, low LSAT, high B, and high RM compared to cheaper homes (top panel), which are more mixed on CRIM, LSAT, and B, and have a medium level of RM. This observation gives an indication of useful predictors and suggests possible binning for some numerical predictors.

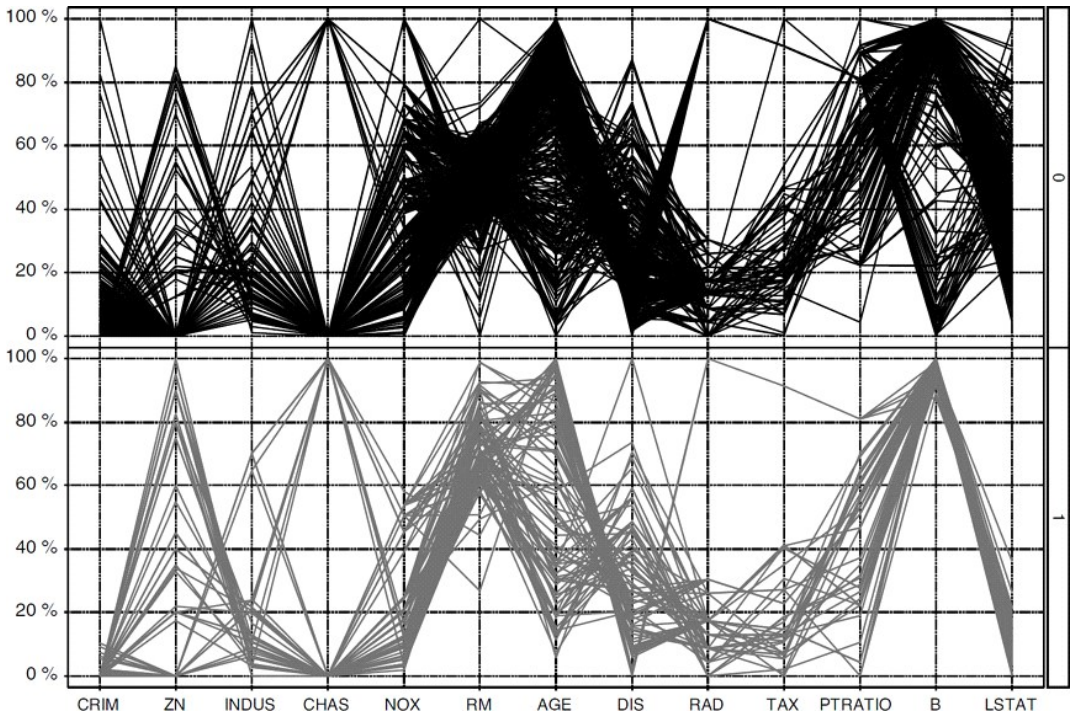

FIGURE 3.11: PARALLEL COORDINATES PLOT FOR BOSTON HOUSING DATA. EACH OF THE VARIABLES (SHOWN ON THE HORIZONTAL AXIS) IS SCALED TO 0—100%. PANELS ARE USED TO DISTINGUISH CAT.MEDV (TOP PANEL = HOMES BELOW $30, 000). CREATED USING SPOTFIRE

Parallel coordinate plots are also useful in unsupervised tasks. They can reveal clusters, outliers, and information overlap across variables. A useful manipulation is to reorder the columns to better reveal observation clusterings. Parallel coordinate plots are not implemented in Excel. However, a free Excel add-in is currently available at http://ibmi.mf.uni-lj.si/ibmi-english/biostat-center/ programje/excel/ParallelCoordinates.xls.

Interactive Visualization

Similar to the interactive nature of the data mining process, interactivity is key to enhancing our ability to gain information from graphical visualization. In the words of Stephen Few (Few, 2009, p. 55), an expert in data visualization,

> We can only learn so much when staring at a static visualization such as a printed graph …If we can't interact with the data …we hit the wall.

By interactive visualization we mean an interface that supports the following principles:

1. Making changes to a plot is *easy, rapid, and reversible*.
2. Multiple concurrent plots can be easily combined and displayed on a single screen.
3. A set of visualizations can be linked such that operations in one display are reflected in the other displays.

Let us consider a few examples where we contrast a static plot generator (e.g., Excel) with an interactive visualization interface.

Histogram rebinning Consider the need to bin a numerical variable using a histogram. A static histogram would require replotting for each new binning choice (in Excel it would require creating the new bins manually). If the user generates multiple plots, then the screen becomes cluttered. If the same plot is recreated, then it is hard to compare to other binning choices. In contrast, an interactive visualization would provide an easy way to change bin width interactively (see, e.g., the slider below the histogram in Figure 3.12), and then the histogram would automatically and rapidly replot as the user changes the bin width.

Aggregation and Zooming Consider a time series forecasting task, given a long series of data. Temporal aggregation at multiple levels is needed for determining short- and long-term patterns. Zooming and panning are used to identify unusual periods. A static plotting software requires the user to create new data columns for each temporal aggregation (e.g., aggregate daily data to obtain weekly aggregates). Zooming and panning in Excel requires manually changing the min and

max values on the axis scale of interest (thereby losing the ability to quickly move between different areas without creating multiple charts). An interactive visualization would provide immediate temporal hierarchies between which the user can easily switch. Zooming would be enabled as a slider near the axis (see, e.g., the sliders on the top left panel in Figure 3.12), thereby allowing direct manipulation and rapid reaction.

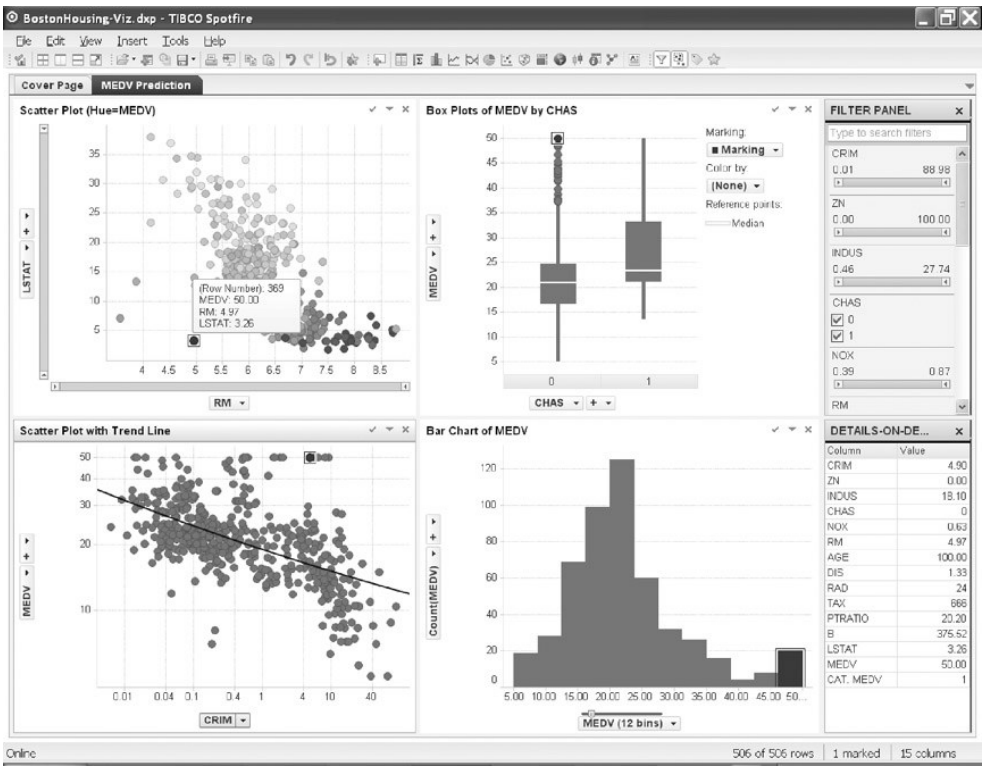

FIGURE 3.12: MULTIPLE INTERLINKED PLOTS IN A SINGLE VIEW (IN SPOTFIRE). NOTE THE MARKED OBSERVATION IN THE TOP LEFT PANEL, WHICH IS ALSO HIGHLIGHTED IN ALL OTHER PLOTS

CombiningMultiple Linked Plots That Fit in a Single Screen To support a classification task, multiple plots are created of the outcome variable versus potential categorical and numerical predictors. These can include side-by-side boxplots, color-coded scatterplots, multipanel bar charts, and the like. The user wants to detect multidimensional relationships (and identify outliers) by selecting a certain subset of the data (e.g., a single category of some variable) and locating the observations on the other plots. In a static interface, the user would have to manually organize the plots of interest and resize them in order to fit within a single screen. A static interface would usually not support interplot linkage, and even if so, the entire set of plots would have to be regenerated each time that a selection is made. In contrast, an interactive visualization would provide an easy way to automatically organize and resize the set of plots to fit within a screen. Linking the set of plots would be easy, and in response to the users selection on one plot, the appropriate selection would be automatically highlighted in the other plots (see example in Figure 3.12).

In earlier sections we used plots to illustrate the advantages of visualizations because "a picture is worth a thousand words." The advantages of an interactive visualization are even greater. As Ben Shneiderman, a well-known researcher in information visualization and interfaces, notes:

> A picture is worth a thousand words. An interface is worth a thousand pictures.

Interactive Visualization Software Some added features such as color, shape, and size are often available in software that produces static plots, while others (multiple panels, hierarchies, labels) are only available in more advanced visualization tools. Even when a feature is available (e.g., color), the ease of applying it to a plot can widely vary. For example, incorporating color into an Excel scatterplot is a daunting task.[2] Plot manipulation possibilities (e.g., zooming, filtering, and aggregation) and ease of implementation are also quite limited in standard "static plot" software.

Although we do not intend to provide a market survey of interactive visualization tools, we do mention a few prominent packages. Spotfire (http://spotfire.tibco.com) and Tableau (www.tableausoftware.com) are two dedicated data visualization tools (several of the plots in this chapter were created using Spotfire). They both provide a high level of interactivity, can

support large datasets, and produce high-quality plots that are also easy to export. JMP by SAS (www.jmp.com) is a "statistical discovery" software that also has strong interactive visualization capabilities. All three offer free trial versions. Finally, we mention Many Eyes by IBM (http://manyeyes.alphaworks.ibm.com/manyeyes) that allows uploading your data and visualizing it via different interactive visualizations.

[2]See http://blog.bzst.com/2009/08/creating-color-coded-scatterplots- in.html.

3.5 Specialized Visualizations

In this section we mention a few specialized visualizations that are able to capture data structures beyond the standard time series and cross-sectional structures— special types of relationships that are usually hard to capture with ordinary plots. In particular, we address hierarchical data, network data, and geographical data— three types of data that are becoming more available.

Visualizing Networked Data

With the explosion of social and product network data, network analysis has become a hot topic. Examples of social networks include networks of sellers and buyers on eBay and networks of people on Facebook. An example of a product network is the network of products on Amazon (linked through the recommendation system). Network data visualization is available in various network-specialized software, and also in general-purpose software.

A network diagram consists of actors and relations between them. "Nodes" are the actors (e.g., people in a social network or products in a product network) and represented by circles. "Edges" are the relations between nodes and are represented by lines connecting nodes. For example, in a social network such as Facebook, we can construct a list of users (nodes) and all the pairwise relations (edges) between users who are "Friends" Alternatively, we can define edges as a posting that one user posts on another user's Facebook page. In this setup we might have more than a single edge between two nodes. Networks can also have nodes of multiple types. A common structure is networks with two types of nodes. An example of a two-type node network is shown in Figure 3.13, where we see a set of transactions between a network of sellers and buyers on the online auction site www.eBay.com [the data are for auctions selling Swarowski beads and took place during a period of several months; from Jank and Yahav (2010)]. The circles on the left side represent sellers and on the right side buyers. Circle size represents the number of transactions that the node (seller or buyer) was involved in within this network. Line width represents the number of auctions that the bidder-seller pair interacted in (in this case we use arrows to denote the directional relationship from buyer to seller). We can see that this marketplace is dominated by three or four high-volume sellers. We can also see that many buyers interact with a single seller. The market structures for many individual products could be reviewed quickly in this way. Network providers could use the information, for example, to identify possible partnerships to explore with sellers.

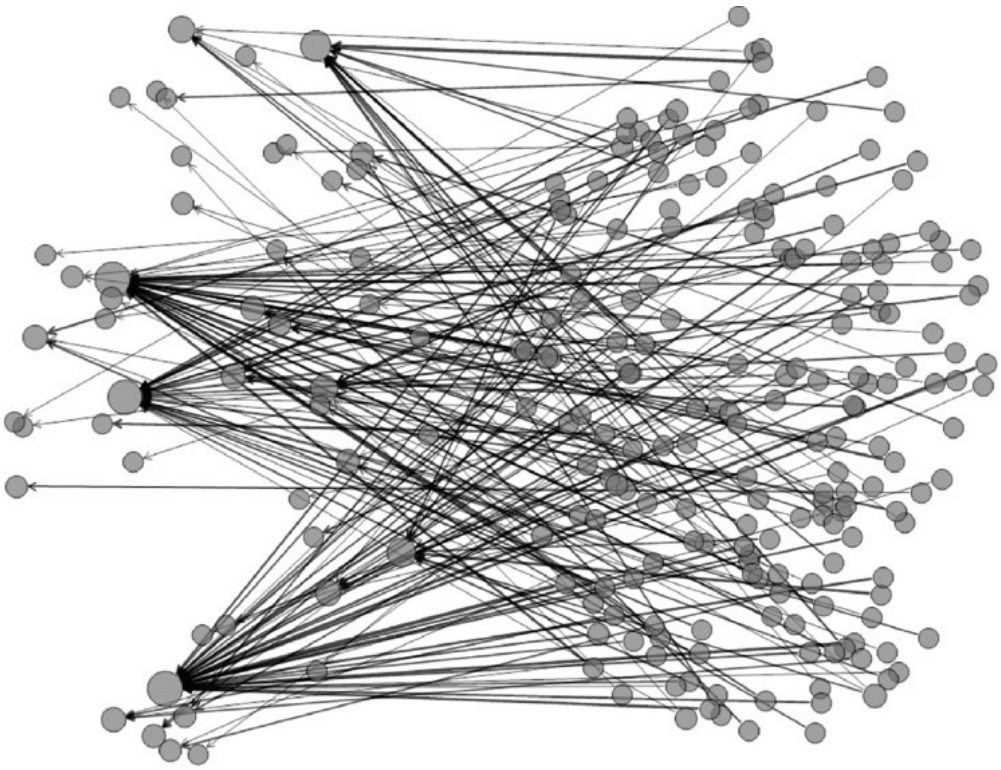

FIGURE 3.13: NETWORK GRAPH OF EBAY SELLERS (LEFT SIDE) AND BUYERS (RIGHT SIDE) OF SWAROSKI BEADS. CIRCLE SIZE REPRESENTS THE NODE'S NUMBER OF TRANSACTIONS. LINE WIDTH REPRESENTS THE NUMBER OF TRANSACTIONS BETWEEN THAT PAIR OF SELLER-BUYER (CREATED WITH SPOTFIRE)

Figure 3.13 was produced using Spotfire's network visualization. An Excel-based tool is NodeXL (http://nodexl.codeplex.com), which is a template for Excel 2007 that allows entering a network edge list. The graph's appearance can be customized and various interactive features are available such as zooming, scaling and panning the graph, dynamically filtering nodes and edges, altering the graph's layout, finding clusters of related nodes, and calculating graph metrics. Networks can be imported from and exported to a variety of data formats, and built-in connections for getting networks from Twitter, Flickr, and your local e-mail are provided.

Network graphs can be potentially useful in the context of association rules (see Chapter 13). For example, consider a case of mining a dataset of consumers' grocery purchases to learn which items are purchased together ("what goes with what"). A network can be constructed with items as nodes and edges connecting items that were purchased together. After a set of rules is generated by the data mining algorithm (which often contains an excessive number of rules, many of which are unimportant), the network graph can help visualize different rules for the purpose of choosing the interesting ones. For example, a popular "beer and diapers" combination would appear in the network graph as a pair of nodes with very high connectivity. An item that is almost always purchased regardless of other items (e.g., milk) would appear as a very large node with high connectivity to all other nodes.

Visualizing Hierarchical Data: Treemaps

We discussed hierarchical data and the exploration of data at different hierarchy levels in the context of plot manipulations. *Treemaps* are useful visualizations for exploring large data sets that are hierarchically structured (tree structured). They enable exploration of various dimensions of the data while maintaining the data's hierarchical nature. An example is shown in Figure 3.14, which displays a large set of eBay auctions, hierarchically ordered by item category, subcategory, and brand. The levels in the hierarchy of the treemap are visualized as rectangles containing subrectangles. Categorical variables can be included in the display by using hue. Numerical variables can be included via rectangle size and color intensity (ordering of the rectangles is sometimes used to reinforce size). In Figure 3.14 size is used to represent the average closing price (which reflects item value), and color intensity represents the percent of sellers with negative feedback (a negative seller feedback indicates buyer dissatisfaction in past transactions and often indicative of fraudulent seller behavior). Consider the task of classifying ongoing auctions in terms of a fraudulent outcome. From the treemap we see that the highest proportion of sellers with negative ratings (black) is concentrated in expensive item auctions (Rolex and Cartier wristwatches).

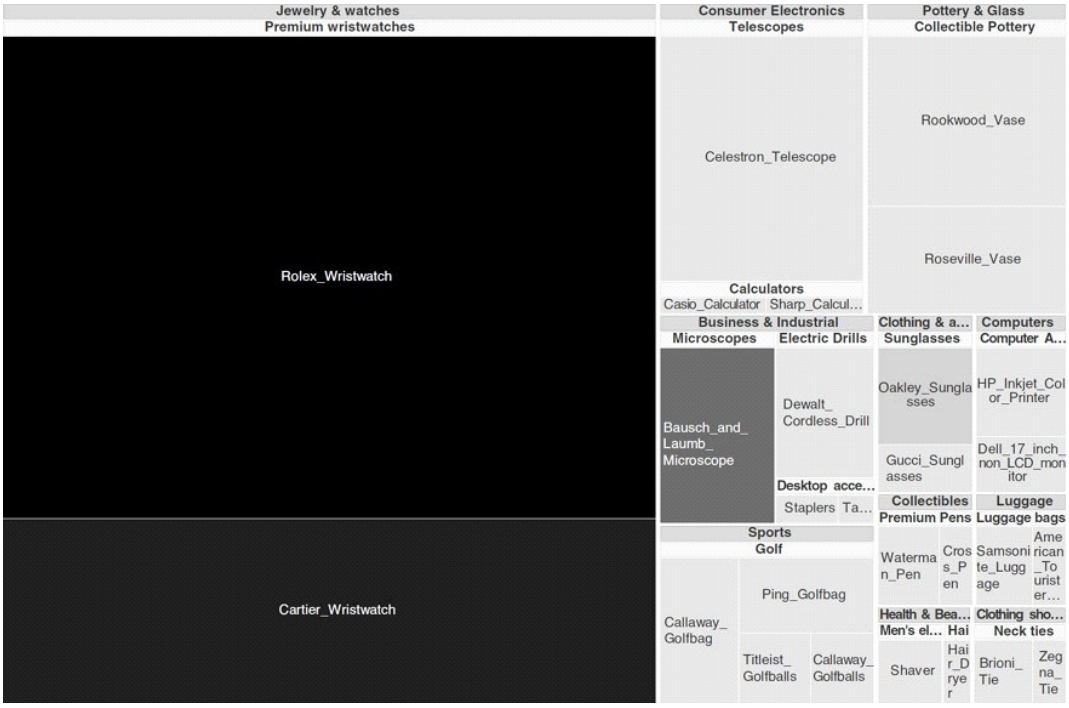

FIGURE 3.14: TREEMAP SHOWING NEARLY 11, 000 EBAY AUCTIONS, ORGANIZED BY ITEM CATEGORY, SUBCATEGORY, AND BRAND. RECTANGLE SIZE REPRESENTS AVERAGE CLOSING PRICE (REFLECTING ITEM VALUE). SHADE REPRESENTS % OF SELLERS WITH NEGATIVE FEEDBACK (DARKER = HIGHER %)

Ideally, treemaps should be explored interactively, zooming to different levels of the hierarchy. An interactive online application of treemaps is "Map of the Market" by Smart-Money (www.smartmoney.com/map-of-the-market), which displays stock market information in an interactive treemap display.

A free treemap add-in for Excel was developed by Microsoft research and is available at http://research.microsoft.com/apps/dp/dl/downloads.aspx (search for "Treemapper").

Visualizing Geographical Data: Map Charts

With the growing availability of location data, many datasets used for data mining now include geographical information. Zip codes are one example of a categorical variable with many categories, where creating meaningful variables for analysis is not straightforward. Plotting the data on a geographical map can often reveal patterns that are otherwise harder to identify. A map chart uses a geographical map as its background, and then color, hue, and other features can be used to include categorical or numerical variables. Besides specialized mapping software, maps are now becoming part of general-purpose software. Figure 3.15 shows two world maps (created with Spotfire), comparing countries' "well-being" (according to a 2006 Gallup survey) in the top map to gross domestic product (GDP) in the bottom map. A darker shade means higher value (white areas are missing data).

FIGURE 3.15: WORLD MAPS COMPARING "WELL-BEING" TO GDP. (TOP) SHADING BY AVERAGE "GLOBAL WELL-BEING" SCORE OF COUNTRY (DARKER CORRESPONDS TO HIGHER SCORE OR LEVEL). (BOTTOM) SHADING ACCORDING TO GDP. DATA FROM VEENHOVEN'S WORLD DATABASE OF HAPPINESS

3.6 Summary of Major Visualizations and Operations, According to Data Mining GOAL

Prediction

- Plot outcome on the *y* axis of vertical boxplots, vertical bar charts, and scatterplots.
- Study relation of outcome to categorical predictors via side-by-side box-plots, bar charts, and multiple panels.
- Study relation of outcome to numerical predictors via scatterplots.
- Use distribution plots (boxplot, histogram) for determining needed transformations of the outcome variable (and/or numerical predictors).
- Examine scatterplots with added color/panels/size to determine the need for interaction terms.
- Use various aggregation levels and zooming to determine areas of the data with different behavior, and to evaluate the level of global versus local patterns.

Classification

- Study relation of outcome to categorical predictors using bar charts with the outcome on the *y* axis.
- Study relation of outcome to pairs of numerical predictors via color-coded scatterplots (color denotes the outcome).
- Study relation of outcome to numerical predictors via side-by-side box-plots: Plot boxplots of a numerical variable by outcome. Create similar displays for each numerical predictor. The most separable boxes indicate potentially useful predictors.
- Use color to represent the outcome variable on a parallel coordinate plot.
- Use distribution plots (boxplot, histogram) for determining needed transformations of the outcome variable.
- Examine scatterplots with added color/panels/size to determine the need for interaction terms.

- Use various aggregation levels and zooming to determine areas of the data with different behavior, and to evaluate the level of global versus local patterns.

Time Series Forecasting

- Create line graphs at different temporal aggregations to determine types of patterns.
- Use zooming and panning to examine various shorter periods of the series to determine areas of the data with different behavior.
- Use various aggregation levels to identify global and local patterns.
- Identify missing values in the series (that require handling).
- Overlay trend lines of different types to determine adequate modeling choices.

Unsupervised Learning

- Create scatterplot matrices to identify pairwise relationships and clustering of observations.
- Use heatmaps to examine the correlation table.
- Use various aggregation levels and zooming to determine areas of the data with different behavior.
- Generate a parallel coordinate plot to identify clusters of observations.

Problems

3.1 **Shipments of Household Appliances: Line Graphs**. The file ApplianceShipments.xls contains the series of quarterly shipments (in million $) of U.S. household appliances between 1985 and 1989 (data courtesy of Ken Black).

a. Create a well-formatted time plot of the data using Excel.

b. Does there appear to be a quarterly pattern? For a closer view of the patterns, zoom in to the range of 3500—5000 on the *y* axis.

c. Create four separate lines for Q1, Q2, Q3, and Q4, using Excel. In each, plot a line graph. In Excel, order the data by Q1, Q2, Q3, Q4 (alphabetical sorting will work), and plot them as separate series on the line graph. Zoom in to the range of 3500—5000 on the *y* axis. Does there appear to be a difference between quarters?

d. Using Excel, create a line graph of the series at a yearly aggregated level (i.e., the total shipments in each year).

e. Re-create the above plots using an interactive visualization tool. Make sure to enter the quarter information in a format that is recognized by the software as a date.

f. Compare the two processes of generating the line graphs in terms of the effort as well as the quality of the resulting plots. What are the advantages of each?

3.2 **Sales of Riding Mowers: Scatterplots**. A company that manufactures riding mowers wants to identify the best sales prospects for an intensive sales campaign. In particular, the manufacturer is interested in classifying households as prospective owners or nonowners on the basis of Income (in $1000s) and Lot Size (in 1000 ft^2). The marketing expert looked at a random sample of 24 households, included in the file RidingMowers.xls.

a. Using Excel, create a scatterplot of Lot Size vs. Income, color coded by the outcome variable owner/nonowner. Make sure to obtain a well-formatted plot (remove excessive background and gridlines; create legible labels and a legend, etc.). The result should be similar to Figure 9.2. *Hint:* First sort the data by the outcome variable, and then plot the data for each category as separate series.

b. Create the same plot, this time using an interactive visualization tool.

c. Compare the two processes of generating the plot in terms of the effort as well as the quality of the resulting plots. What are the advantages of each?

3.3 **Laptop Sales at a London Computer Chain: Bar Charts and Boxplots**. The file LaptopSalesJanuary2008.xls contains data for all sales of laptops at a computer chain in London in January 2008. This is a subset of the full dataset that includes data for the entire year.

a. Create a bar chart, showing the average retail price by store. Which store has the highest average? Which has the lowest?

b. To better compare retail prices across stores, create side-by-side boxplots of retail price by store. Now compare the prices in the two stores above. Do you see a difference between their price distributions? Explain.

3.4 **Laptop Sales at a London Computer Chain: Interactive Visualization**. *The next exercises are designed for use with an interactive visualization tool. The file LaptopSales.txt is a comma-separated file with nearly 300, 000 rows. ENBIS (the European Network for Business and Industrial Statistics) provided these data as part of a contest organized in the fall of 2009*.

Scenario: You are a new analyst for Acell, a company selling laptops. You have been provided with data about products and sales. Your task is to help the company to plan product strategy and pricing policies that will maximize Acell's projected revenues in 2009. Using an interactive visualization tool, answer the following questions.

a. Price Questions
   i. At what prices are the laptops actually selling?
   ii. Does price change with time? *(Hint:* Make sure that the date column is recognized as such. The software should then enable different temporal aggregation choices, e.g., plotting the data by weekly or monthly aggregates, or even by day of week.)
   iii. Are prices consistent across retail outlets?
   iv. How does price change with configuration?

b. Location Questions
   i. Where are the stores and customers located?
   ii. Which stores are selling the most?
   iii. How far would customers travel to buy a laptop?
      - *Hint 1:* you should be able to aggregate the data, for example, plot the sum or average of the prices.
      - *Hint 2:* Use the coordinated highlighting between multiple visualizations in the same page, for example, select a store in one view to see the matching customers in another visualization.
      - *Hint 3:* Explore the use of filters to see differences. Make sure to filter in the zoomed out view. For example, try to use a "store location" slider as an alternative way to dynamically compare store locations. This is especially useful for spotting outlier patterns if there are many store locations to compare.
   iv. Try an alternative way of looking at how far customers traveled. Do this by creating a new data column that computes the distance between customer and store.

c. Revenue Questions
   i. How do the sales volume in each store relate to Acell's revenues?
   ii. How does this depend on the configuration?

d. Configuration Questions
   i. What are the details of each configuration? How does this relate to price?
   ii. Do all stores sell all configurations?