Chapter 13

Dealing in Graphic Detail

In This Chapter

- Using familiar graphs in familiar ways
- Using familiar graphs in unfamiliar ways
- Adding new graphs to your bag of tricks

You may be new to data mining, but you already know some important tools of the trade. Graphs like bar charts, histograms, and scatterplots are important data-mining tools. Data miners use these conventional graphs in conventional ways, and in unconventional ways, too! And now that you're a data miner, you can expand your repertoire with special graphs that help you pack more information on a page (without losing main ideas), spot common patterns, or evaluate predictive models.

This chapter introduces you to the data miner's arsenal of graphs and graphing tools. You'll find that graphs are one of the easiest ways to get started in data mining, especially since data miners often use the sorts of graphs (or variations of those graphs) you've probably already used elsewhere.

Starting Simple

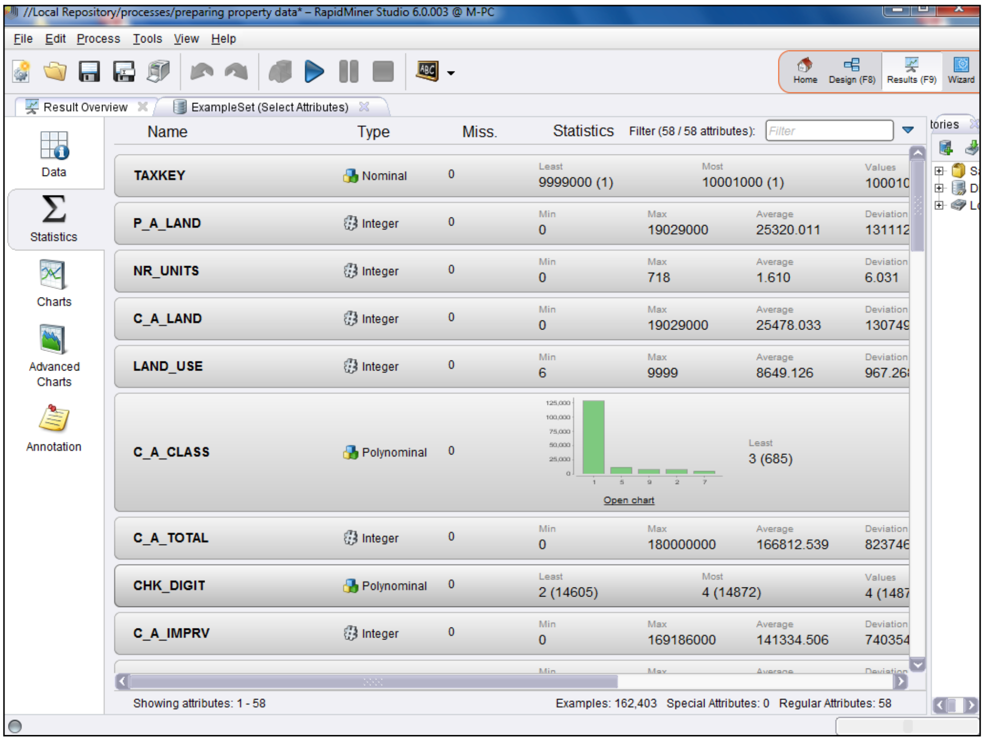
All data miners use graphs, and all data-mining applications offer some graphics capability. Some data-mining applications offer only graphs that you might remember from your elementary school days, like bar charts and scatterplots. That's because these simple graphs are the ones that data miners use most often.

Eyeballing variables with bar charts and histograms

A basic part of the data-understanding phase of the data-mining process (refer to Chapter 5 for much more about this process) is investigating variables one at a time, reviewing their distributions, and checking for obvious data quality issues. Bar charts and histograms are visual summaries that make it easy and quick to understand variable distributions.

The two chart types are very similar. If the variable is categorical, use a bar chart; it will have one bar for each category, and the height of the bar shows the frequency of each category. If the variable is continuous, use a histogram. In the histogram, each bar represents a range of values for the variable.

Your data-mining application may make it very easy to get these charts. They are often included in the output of general-purpose data summary tools, like the example shown in Figure 13-1.



Filter (58 / 58 attributes): Filter

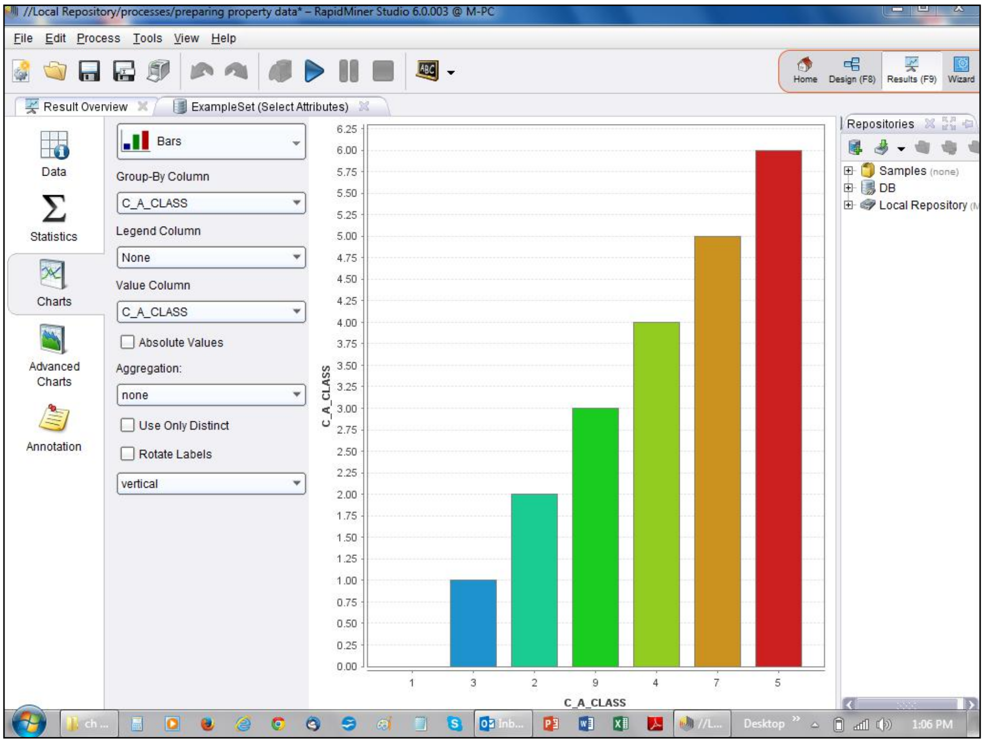
Name	Type	Miss.	Statistics			
TAXKEY	Nominal	0	Least 9999000 (1)	Most 10001000 (1)		Values 100010
P_A_LAND	Integer	0	Min 0	Max 19029000	Average 25320.011	Deviation 131112
NR_UNITS	Integer	0	Min 0	Max 718	Average 1.610	Deviation 6.031
C_A_LAND	Integer	0	Min 0	Max 19029000	Average 25478.033	Deviation 130749
LAND_USE	Integer	0	Min 6	Max 9999	Average 8649.126	Deviation 967.26
C_A_CLASS	Polynominal	0			Least 3 (685)	
C_A_TOTAL	Integer	0	Min 0	Max 180000000	Average 166812.539	Deviation 823746
CHK_DIGIT	Polynominal	0	Least 2 (14605)	Most 4 (14872)		Values 4 (1487
C_A_IMPRV	Integer	0	Min 0	Max 169186000	Average 141334.506	Deviation 740354

Showing attributes: 1 - 58

Examples: 162,403 Special Attributes: 0 Regular Attributes: 58

Figure 13-1: Bar chart included in the output of a data summary tool in RapidMiner.

But it isn't always simple to get the chart you want. Look closely at Figure 13-1, and you'll see the phrase *Open chart* beneath the bar chart. Clicking this link opens a chart editor. You'd expect to see a chart that's identical to the one in the data summary open in the editor, right? Figure 13-2 shows the chart editor as it looks when opened this way. Not identical! You'll have to fuss with setup (see Figure 13-3) to get back to the same point.



Bars

Group-By Column: C_A_CLASS

Legend Column: None

Value Column: C_A_CLASS

□ Absolute Values

Aggregation: none

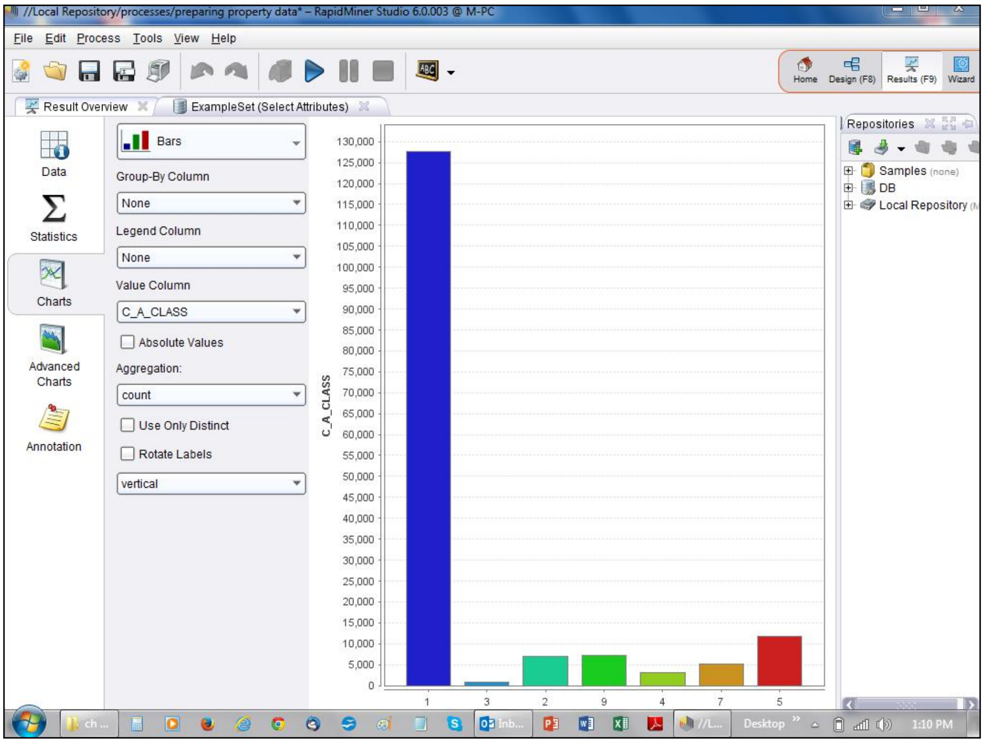
□ Use Only Distinct

□ Rotate Labels

vertical

Figure 13-2: Charts look different when opened in a chart editor!

But this chart editor offers value in other ways. It gives you more options, such as creating more sophisticated chart structure (Figure 13-4 shows an editor that allows complex graph structure) or controlling cosmetic elements like color. Charts editors also provide pathways to export graphs to use in your reports or presentations.



Bars

Group-By Column: None

Legend Column: None

Value Column: C_A_CLASS

☐ Absolute Values

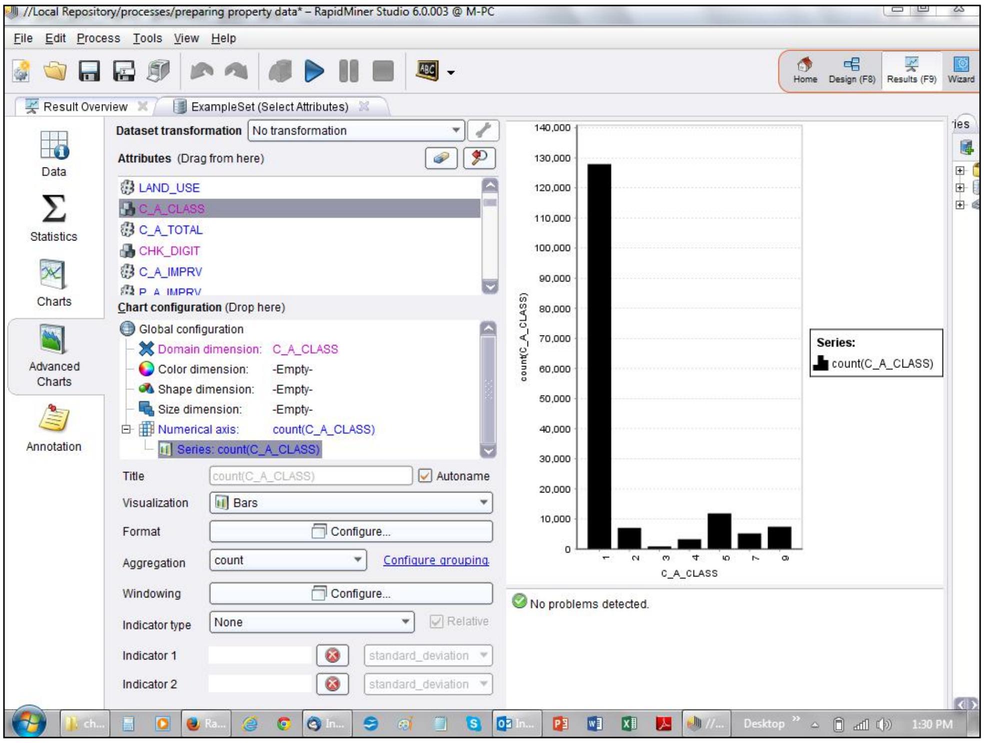
Aggregation: count

☐ Use Only Distinct

☐ Rotate Labels

vertical

Figure 13-3: Correcting the setup to get a basic bar chart.



Dataset transformation: No transformation

Attributes (Drag from here)

- LAND_USE
- C_A_CLASS
- C_A_TOTAL
- CHK_DIGIT
- C_A_IMPRV
- P_A_IMPRV

Chart configuration (Drop here)

- Global configuration
  - Domain dimension: C_A_CLASS
  - Color dimension: -Empty-
  - Shape dimension: -Empty-
  - Size dimension: -Empty-
  - Numerical axis: count(C_A_CLASS)
    - Series: count(C_A_CLASS)

Title: count(C_A_CLASS) ☑ Autoname

Visualization: Bars

Format: Configure...

Aggregation: count Configure grouping

Windowing: Configure...

Indicator type: None ☑ Relative

Indicator 1: standard_deviation

Indicator 2: standard_deviation

Series: count(C_A_CLASS)

No problems detected.

Figure 13-4: A complex chart editor.

The complexities of chart setup seen in this section are matters of product design. A data-mining application may make some operations very easy and others remarkably complex, or not possible. No one magic product outshines all others for ease of use, but one may fit your work style better than others. So, before you settle on a product to use, give it a thorough tryout for the kind of work that you need to do.

Relating one variable to another with scatterplots

The first step toward predictive modeling is relating variables to one another. A simple, remarkable tool for that is the scatterplot. It's used to relate one continuous measure to another. Data miners sometimes stretch the rules and use it with categorical variables as well.

The horizontal (*x*) axis of the plot represents values of one variable; the vertical axis (*y*) represents a second variable. You may not have a sense of which variable is independent and which is dependent for every pair of variables. If you do, the independent variable should be on the horizontal axis. Each point on the plot represents the coordinates, the pair of values for the two variables within a single case. (These pairs are sometimes called *xy pairs*).

Find your scatterplot tool (Figure 13-5 shows this tool on the menu of Orange; the location for the tool varies by product) and set up a basic scatterplot tool by selecting two variables to use. The example in Figure 13-6 shows an interactive display; the scatterplot appears immediately. In another tool, you might need additional steps to execute and create the chart.

The scatterplot example in Figure 13-6 relates auto mileage to engine horsepower. Low horsepower is associated with high mileage, and the higher the horsepower, the lower the mileage. You can easily see this pattern in the data. You might notice a shape, not linear but somewhat curved. This could provide hints about what model types to try later.

Data-mining applications often have some interactive features in graph displays. For example, Figure 13-7 shows that hovering your mouse over a point reveals the exact values of the two variables for that point. This is easier than trying to read the values from the axes!

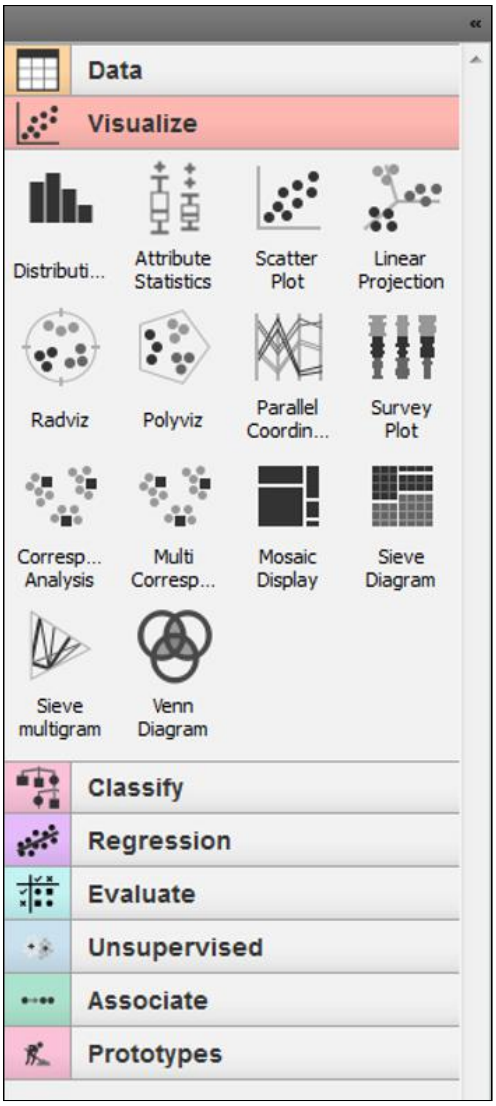
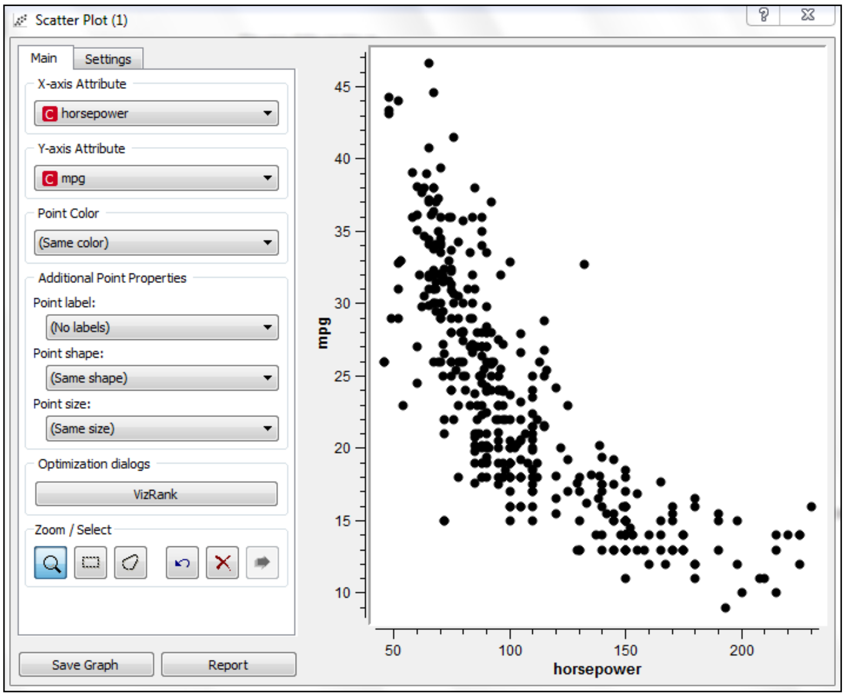


Figure 13-5: Finding the scatterplot tool in Orange.



Scatter Plot (1)

X-axis Attribute: horsepower

Y-axis Attribute: mpg

Point Color: (Same color)

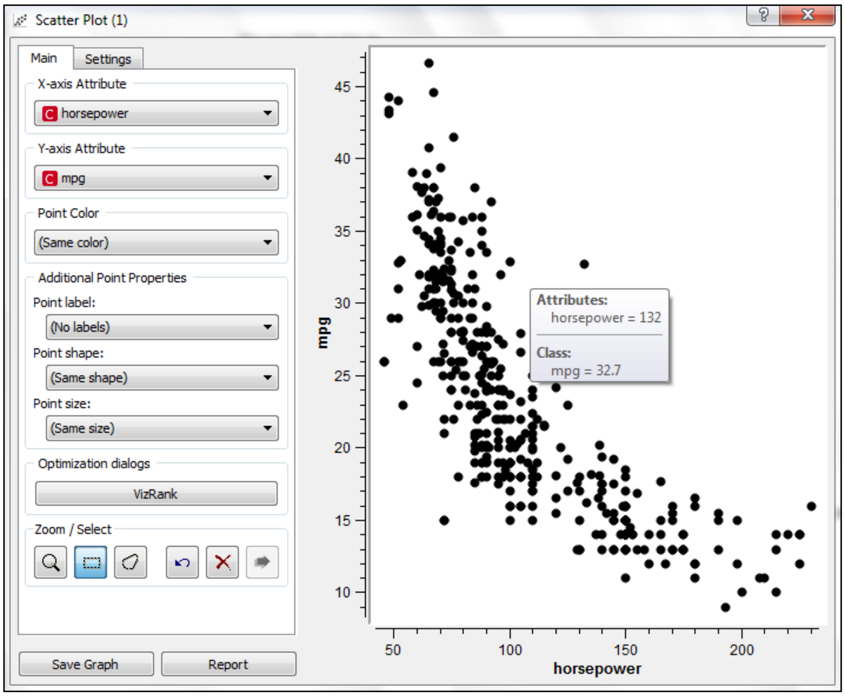
Additional Point Properties

Point label: (No labels)

Point shape: (Same shape)

Point size: (Same size)

Figure 13-6: Scatterplot of auto mileage versus horsepower.



Scatter Plot (1)

X-axis Attribute: horsepower

Y-axis Attribute: mpg

Point Color: (Same color)

Additional Point Properties

Point label: (No labels)

Point shape: (Same shape)

Point size: (Same size)

Attributes:
horsepower = 132

Class:
mpg = 32.7

Figure 13-7: Hovering reveals details.

Just say *No* to pie charts

Data miners often use bar charts, histograms, and scatterplots, but you may never meet one who uses pie charts.

What makes a bar chart so different from a pie chart? Both represent relative frequencies of categories. A bar chart is linear, and viewers visually compare the lengths of the bar. A pie chart is not linear. In a pie chart, the relative frequencies are represented by area on the graph surface.

It's hard for people to accurately compare areas. People are pretty good at comparing lengths (one dimension) and not as good at comparing areas (two dimensions). You don't even want to know what happens with volumes (three dimensions) or exponential scales. How do I know this? Research! (Who did the research? Me, I did it. So there!)

What's worse, the pie chart is round. If people can't accurately compare the areas of rectangles (as I know from research), I might suspect that they're even worse at comparing pie slice–shaped areas.

You'll find no glamour in using fussy charts. Avoid nonlinear representations of any kind. Don't use pies, cute shapes, or nonlinear scales. Don't use three-dimensional bar charts. Keep your graphs simple and informative to get good results and support good business decisions.

Building on Basics

Data miners often take advantage of special features to pack more information into simple charts. Labels, overlays, and interactive selection are hallmarks of data-mining applications, special features that allow you to be more productive.

Making scatterplots say more

Figure 13-6 shows a scatterplot that relates auto mileage to engine horsepower. You can see that mileage drops as horsepower increases. That's a beginning step in understanding the factors that determine mileage.

Mileage decreases as horsepower increases, as seen in Figure 13-8 (this is the same scatterplot shown in Figure 13-6, exported as you might to use it in a report or presentation). Mileage increases with time, as you can see in Figure 13-9, a scatterplot of mileage versus model year. It would be helpful to get these two ideas into one graph.

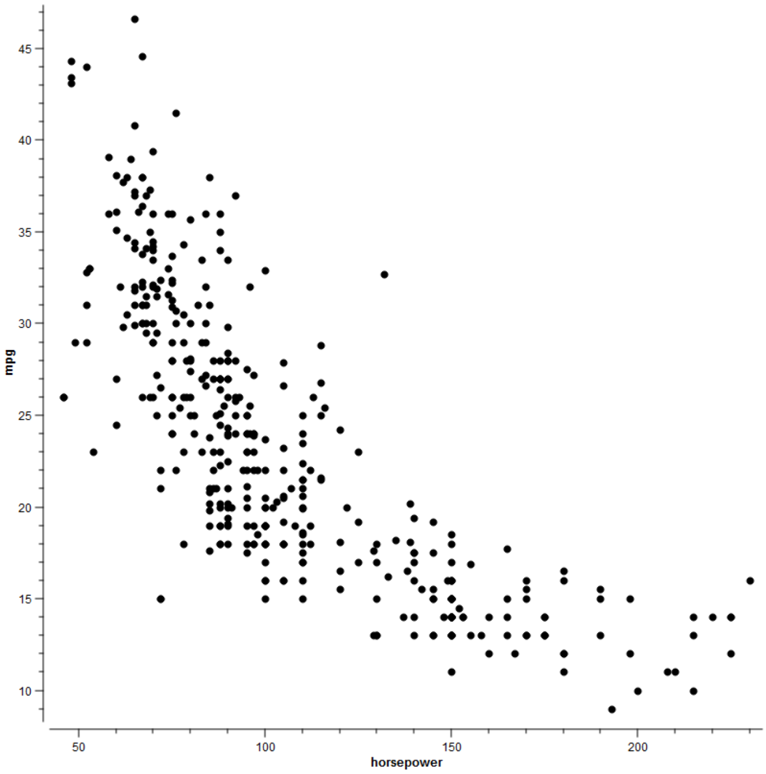


Figure 13-8: A scatterplot of mileage versus horsepower.

Common data-mining approaches for integrating more than two variables in a graph include

- **Labels:** Labels are values of a string or categorical variable that have been superimposed on the scatterplot. Figure 13-10 shows a scatterplot labeled with the model year of the car. (Datasets with many points or long labels can make these charts unreadable, though! The solution is to use only a sample of the data. Setup for this kind of sampling is shown in Figure 13-11.)
- **Overlays:** With overlays, values of a categorical variable define the points' shape or color. Figure 13-12 shows the setup for a scatterplot to overlay model year on the mileage-versus-horsepower scatterplot, and the exported overlay scatterplot appears in Figure 13-13. It may be easier to read color overlays than point shape overlays. The setup is usually much the same.

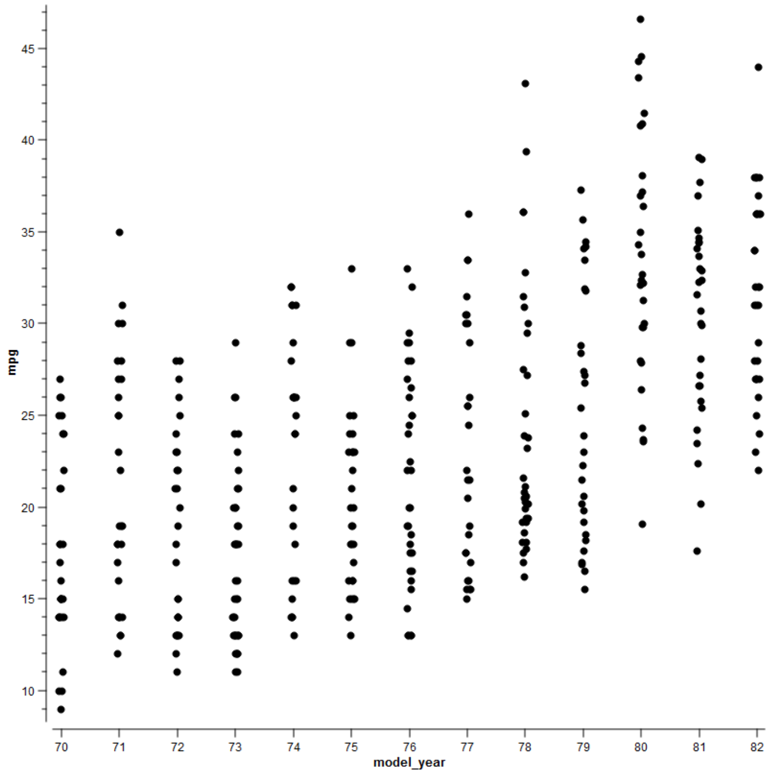


Figure 13-9: A scatterplot of mileage versus model year.

WARNING!

Another thing to keep in mind with scatterplots: You may have multiple points falling on the very same spot! If so, you may not be able to tell a point for one case from a point for 100 cases. The remedy is to check for an option to make multiple instances visible. Look for point size or *jitter* (moves points slightly off their true locations to make all of them visible) options.

Interacting with scatterplots

Interactive scatterplots are great time-savers for data miners.

Say that you see an interesting group of cases in a graph, and you want to further investigate just those cases. If you're looking at just one or two points, you might get the information you want by hovering, as shown in Figure 13-7, but that's not satisfactory when you are interested in more than a couple of points.

Data selection tools in interactive scatterplots give you more power to select data. Figure 13-14 shows the same graph setup, but with a group of points selected by clicking and dragging the mouse around them. This is not just a visual feature. You can export the selected points as a new dataset (see Figure 13-15). This is very handy and fast!

If the points you need don't fit nicely into a rectangular selection, you have other options. Refer to the Zoom/Select area in Figure 13-14. You can see a button with a rectangle for rectangular selection and another with a roundish shape for free-form selection.

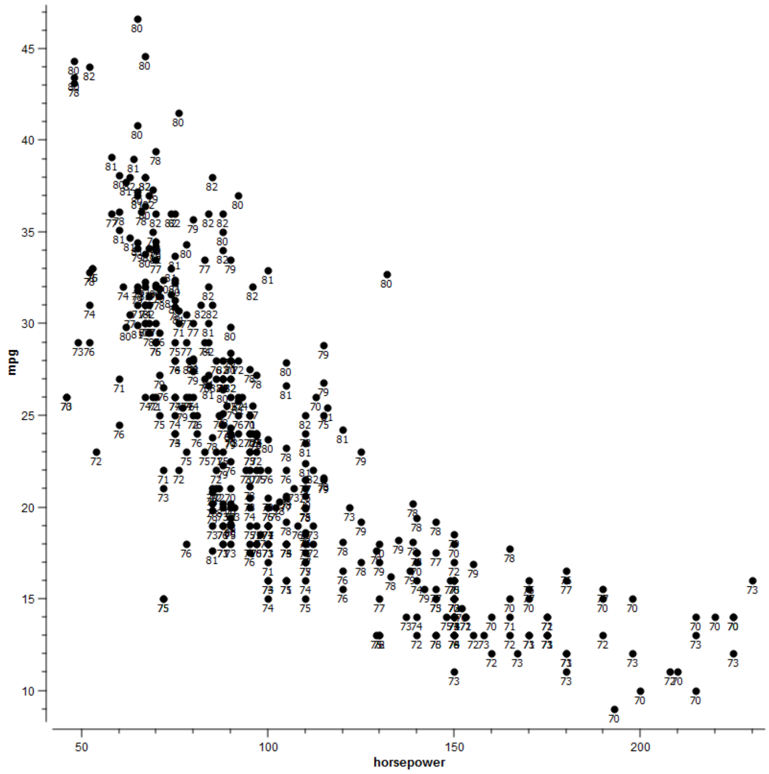
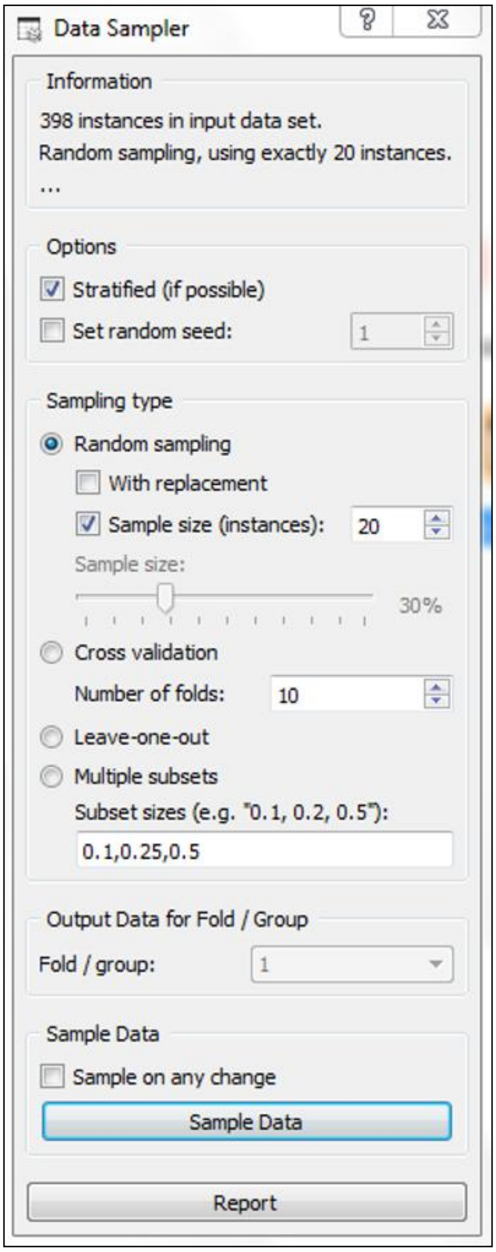


Figure 13-10: A scatterplot with labels.



Data Sampler

Information

398 instances in input data set.

Random sampling, using exactly 20 instances.

...

Options

☑ Stratified (if possible)

□ Set random seed: 1

Sampling type

● Random sampling

□ With replacement

☑ Sample size (instances): 20

Sample size: 30%

○ Cross validation

Number of folds: 10

○ Leave-one-out

○ Multiple subsets

Subset sizes (e.g. "0.1, 0.2, 0.5"): 0.1,0.25,0.5

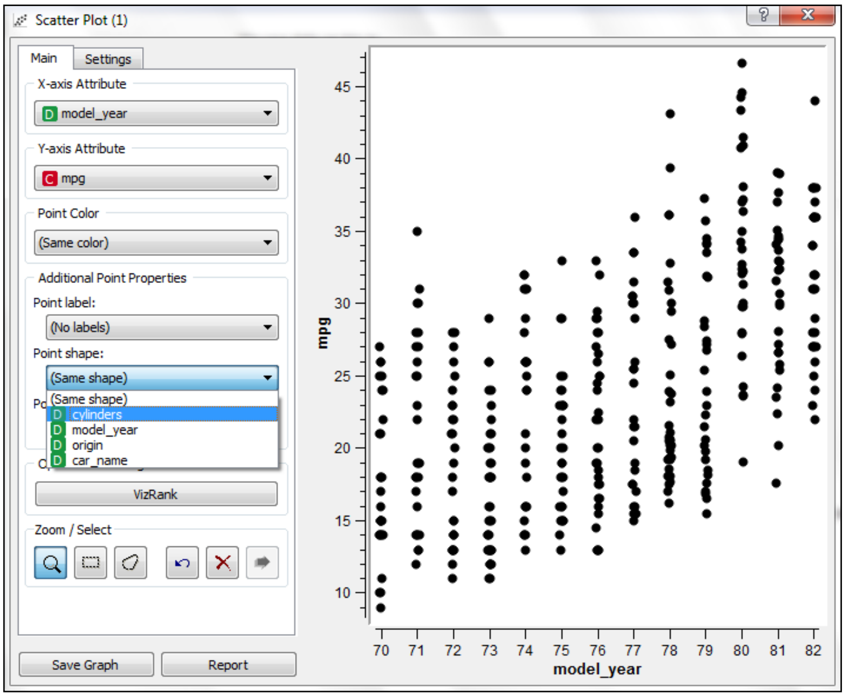
Output Data for Fold / Group

Fold / group: 1

Sample Data

□ Sample on any change

Figure 13-11: Setting up a random sample for a scatterplot.



Scatter Plot (1)

Main | Settings

X-axis Attribute: D model_year

Y-axis Attribute: C mpg

Point Color: (Same color)

Additional Point Properties

Point label: (No labels)

Point shape: (Same shape)

- (Same shape)
- D cylinders
- D model_year
- D origin
- D car_name

Figure 13-12: Setting up an overlay.

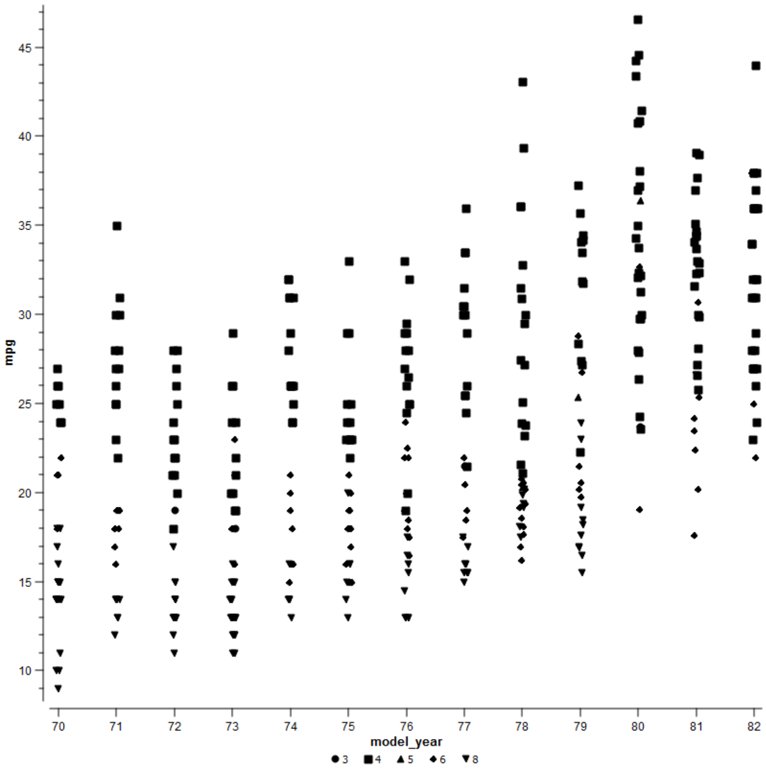
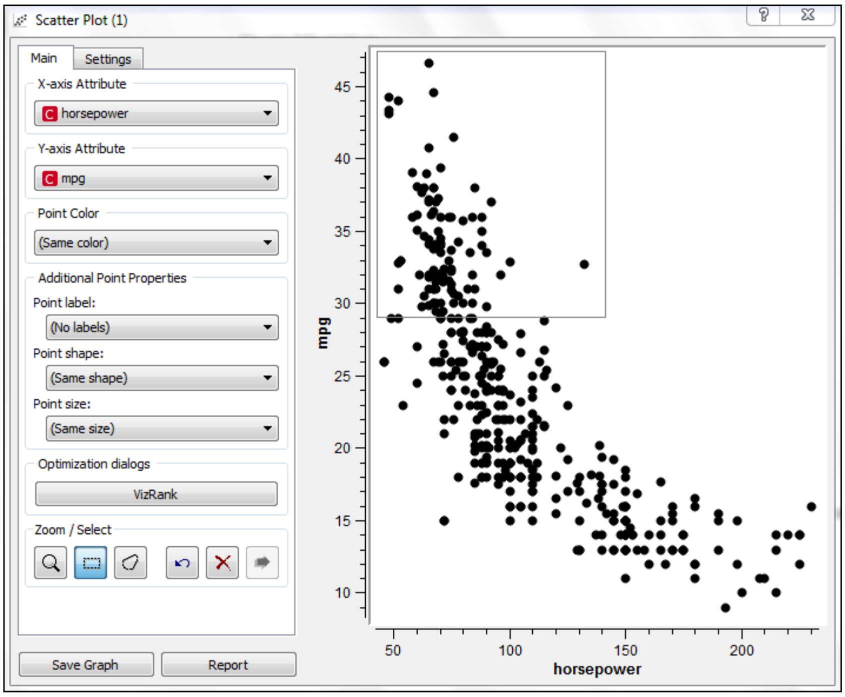


Figure 13-13: Overlay scatterplot using point shapes.



Scatter Plot (1)

Main | Settings

X-axis Attribute: horsepower

Y-axis Attribute: mpg

Point Color: (Same color)

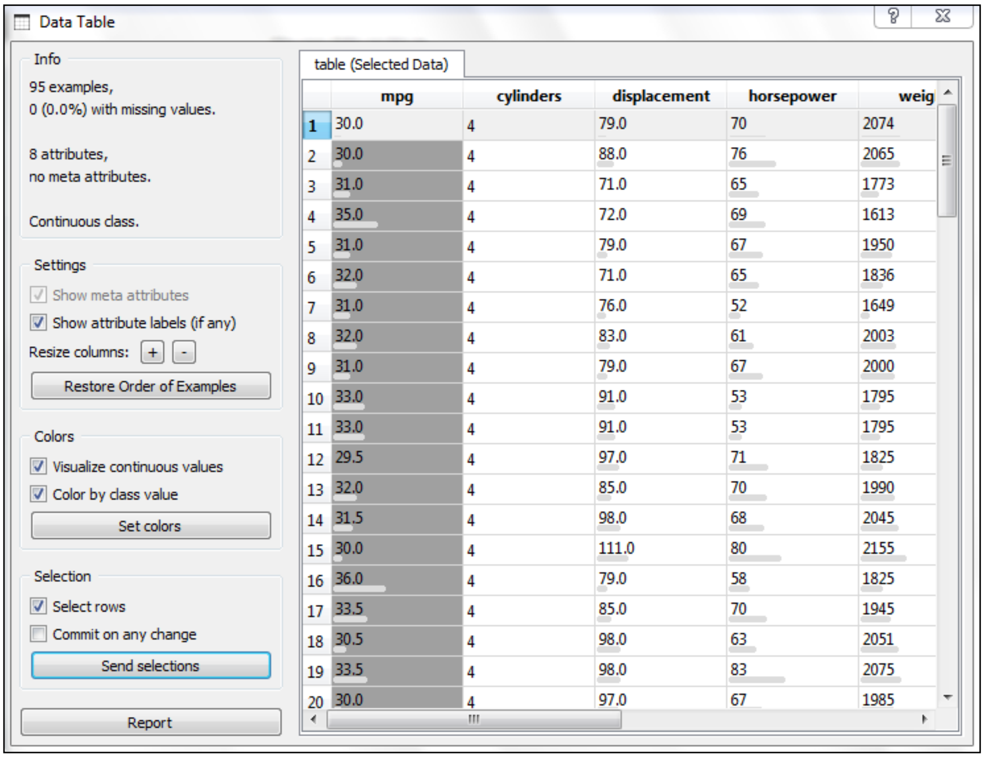
Additional Point Properties

Point label: (No labels)

Point shape: (Same shape)

Point size: (Same size)

Figure 13-14: Rectangular selection.



Data Table

Info

95 examples,
0 (0.0%) with missing values.

8 attributes,
no meta attributes.

Continuous class.

Settings

☑ Show meta attributes

☑ Show attribute labels (if any)

Resize columns: + -

Colors

☑ Visualize continuous values

☑ Color by class value

Selection

☑ Select rows

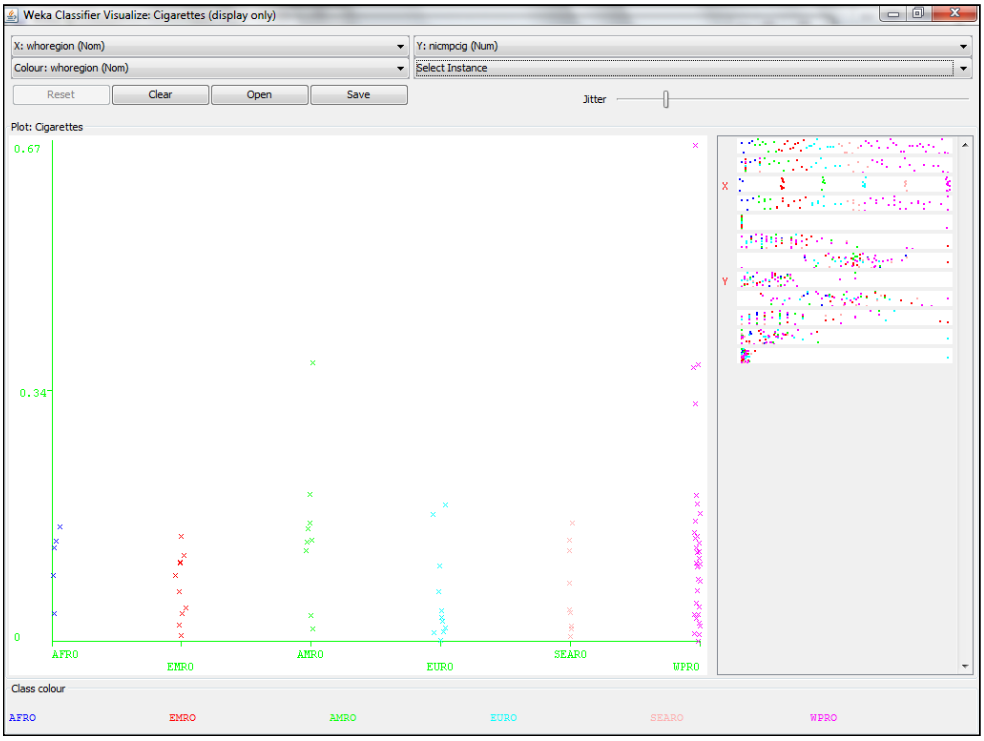
□ Commit on any change

	mpg	cylinders	displacement	horsepower	weig
1	30.0	4	79.0	70	2074
2	30.0	4	88.0	76	2065
3	31.0	4	71.0	65	1773
4	35.0	4	72.0	69	1613
5	31.0	4	79.0	67	1950
6	32.0	4	71.0	65	1836
7	31.0	4	76.0	52	1649
8	32.0	4	83.0	61	2003
9	31.0	4	79.0	67	2000
10	33.0	4	91.0	53	1795
11	33.0	4	91.0	53	1795
12	29.5	4	97.0	71	1825
13	32.0	4	85.0	70	1990
14	31.5	4	98.0	68	2045
15	30.0	4	111.0	80	2155
16	36.0	4	79.0	58	1825
17	33.5	4	85.0	70	1945
18	30.5	4	98.0	63	2051
19	33.5	4	98.0	83	2075
20	30.0	4	97.0	67	1985

Figure 13-15: Dataset created by selection in a graph.

Here's a free-form selection example using data on the nicotine content of cigarettes sold in different parts of the world. This scatterplot (shown in Figure 13-16) shows nicotine per cigarette for samples from the six United Nations regions. (This is a nontraditional use of a scatterplot, because region is not a continuous variable; it's categorical. Data miners often use traditional tools in nontraditional ways.) The points within a region don't fall in a perfect vertical line. Small shifts (jitter) to the left and right are made for readability and appearance only. A few cigarettes have exceptionally high levels of nicotine, and you want to select those cases.

A drop-down menu (see Figure 13-17) offers selection options. Polygon selection lets you mark a free-form area on the scatterplot. To mark, click on the graph to make a starting point, and then click again and again around the group of points you want until you have made the shape you need. (See Figure 13-18.) A right-click indicates that you have completed the selection; this is visible from the highlight on the graph. (See Figure 13-19.)



Weka Classifier Visualize: Cigarettes (display only)

X: whoregion (Nom)

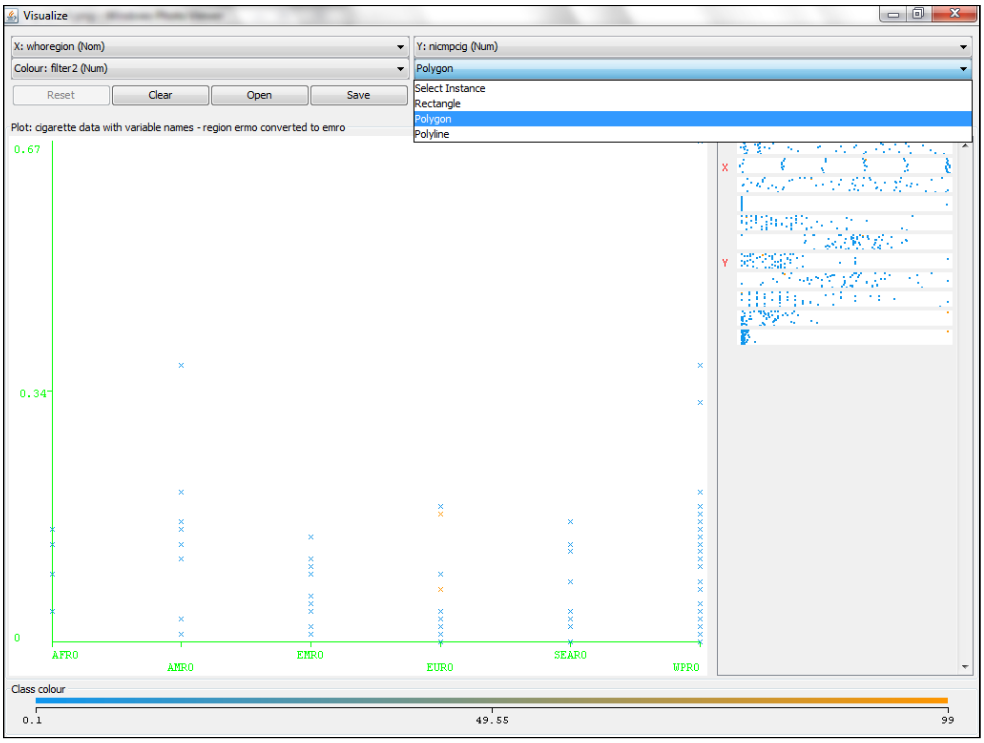
Y: nicmpcig (Num)

Colour: whoregion (Nom)

Select Instance

Plot: Cigarettes

Figure 13-16: Interactive scatterplot in Weka.



Visualize

X: whoregion (Nom)

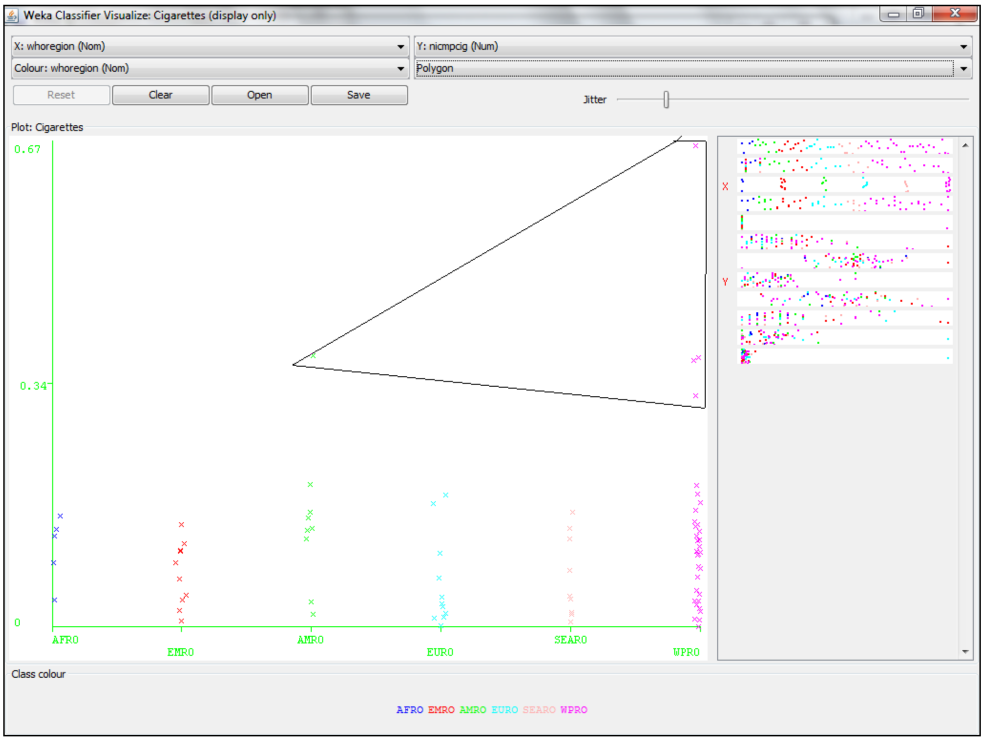
Y: nicmpcig (Num)

Colour: filter2 (Num)

Polygon

Plot: cigarette data with variable names - region ermo converted to emro

Figure 13-17: Choosing polygon selection.



Weka Classifier Visualize: Cigarettes (display only)

X: whoregion (Nom)

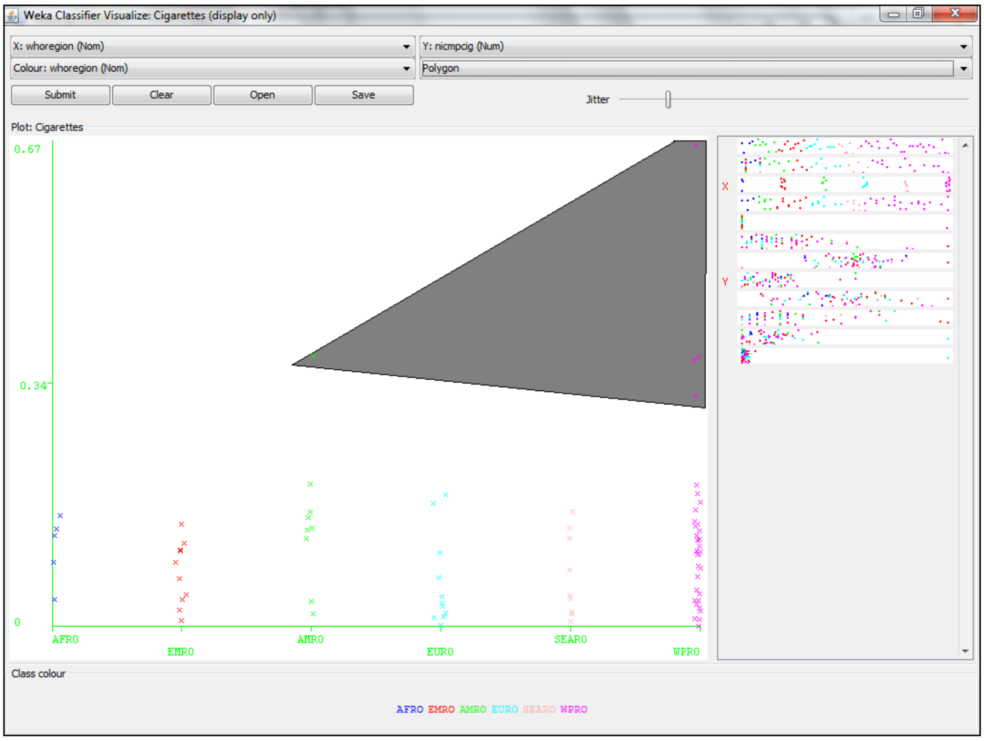
Y: nicmpcig (Num)

Colour: whoregion (Nom)

Polygon

Plot: Cigarettes

Figure 13-18: Clicking points on the graph defines the selected area.



Weka Classifier Visualize: Cigarettes (display only)

X: whoregion (Nom)

Y: nicmpcig (Num)

Colour: whoregion (Nom)

Polygon

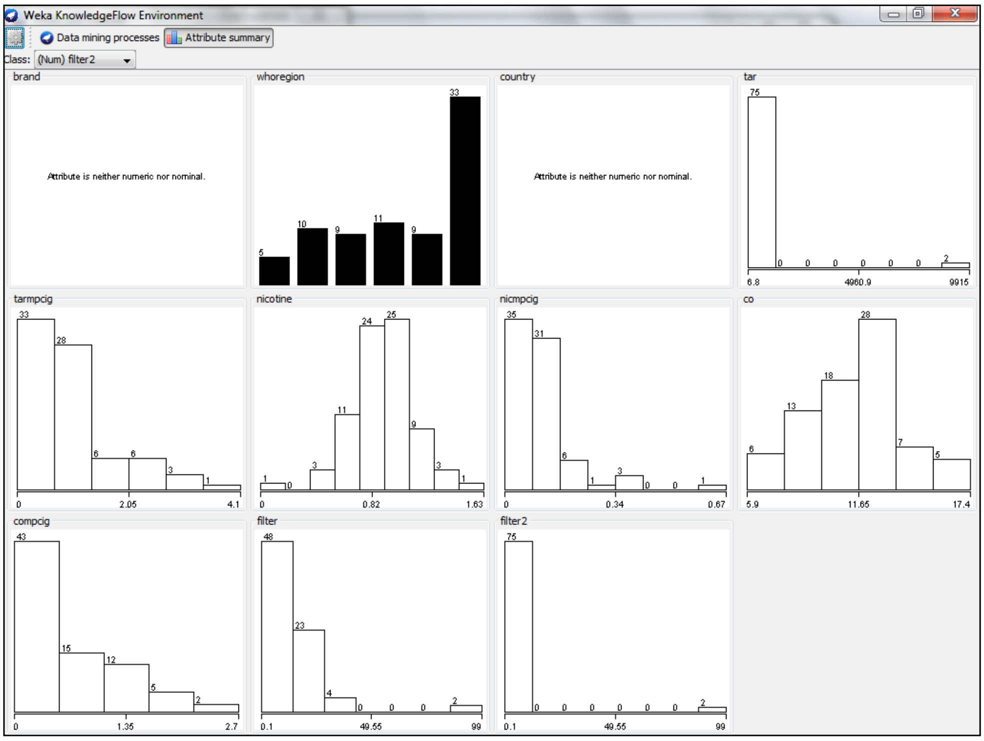
Figure 13-19: Highlighting indicates completed selection.

Working Fast with Graphs Galore

Data miners work fast. One way to improve your productivity is to take full advantage of tools that let you do several things at once.

It's time-consuming (and boring) to set up a number of graphs separately, one at a time. So use these alternatives whenever you can:

- **Data summaries:** Tools that let you quickly ask for summaries of many variables, and get the summaries all at once. Bar charts and histograms are often included in the output. (Refer to Figure 13-1 for an example.)
- **Chart matrix:** The output is a grid (*matrix*) of small bar charts and histograms, so you can review many data distributions quickly. Figures 13-20 and 13-21 show two examples. They were created with different products but are very similar in function and appearance.
- **Scatterplot matrix:** A grid of small scatterplots. Each small graph shows the relationship for a single pair of variables (usually uses continuous variables). You input a list of variables (see Figure 13-22), and the scatterplot matrix shows you every possible pairing.


Weka KnowledgeFlow Environment

Data mining processes | Attribute summary

Class: (Num) filter2

Attribute is neither numeric nor nominal.

Attribute is neither numeric nor nominal.

Figure 13-20: An example chart matrix from Weka.

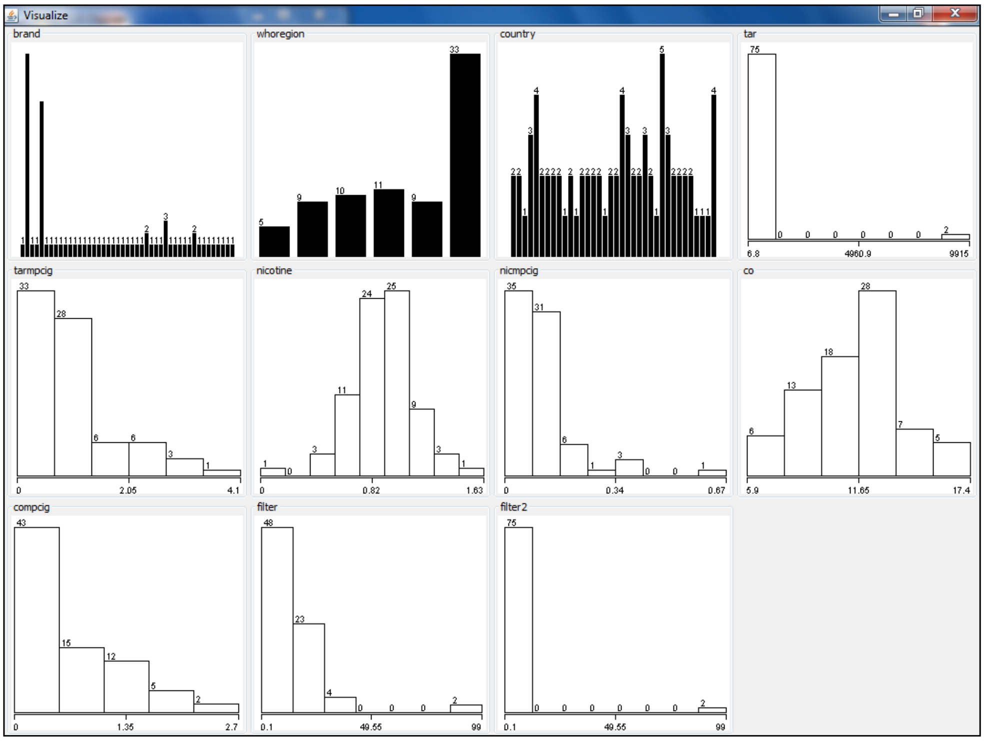
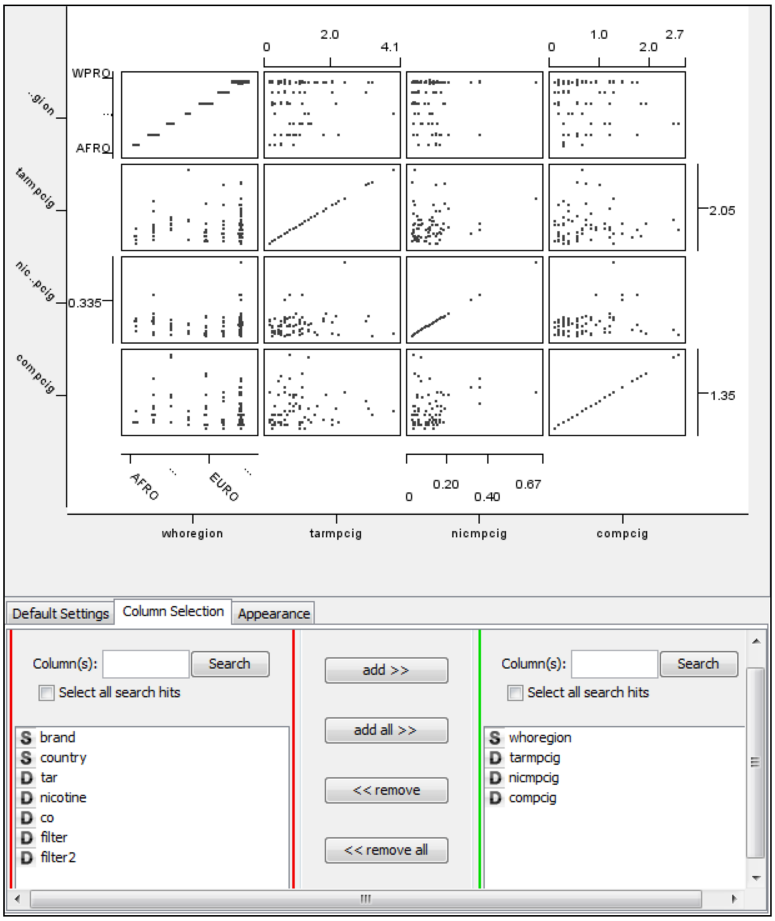

Visualize

Figure 13-21: Another example chart matrix from Weka.



□ Select all search hits

S brand
S country
D tar
D nicotine
D co
D filter
D filter2

□ Select all search hits

S whoregion
D tarmpcig
D nicmpcig
D compcig

Figure 13-22: A scatter-plot matrix.

Extending Your Graphics Range

Because data miners lean heavily on basic graphs, some data-mining applications offer little or nothing more. Others provide a wide range of graph options, from the common to the exotic. It's not necessary to use all of these (read what one expert has to say about this in the nearby sidebar "Putting graphics in context: An interview with Laura Kippen"), but you may benefit by selecting and using a few that suit your own needs.

Data miners often use these graphs:

- **Boxplot (also called *box and whiskers*):** Histograms describe distributions of continuous variables, but have limited value for showing details. A boxplot (see Figure 13-23) is an alternative. The heart of the image is a box; this represents half of the data, taken in the middle of its range. The center of the box is the median value of the variable, and the lower and upper ends of the box represent the 25th- and 75th-percentile levels, respectively. Whiskers extend below and above the box, representing the range of the bulk of the data. Points beyond the whiskers are taken to be *outliers,* highly atypical values (some plots also indicate *extremes,* which are outliers among outliers).

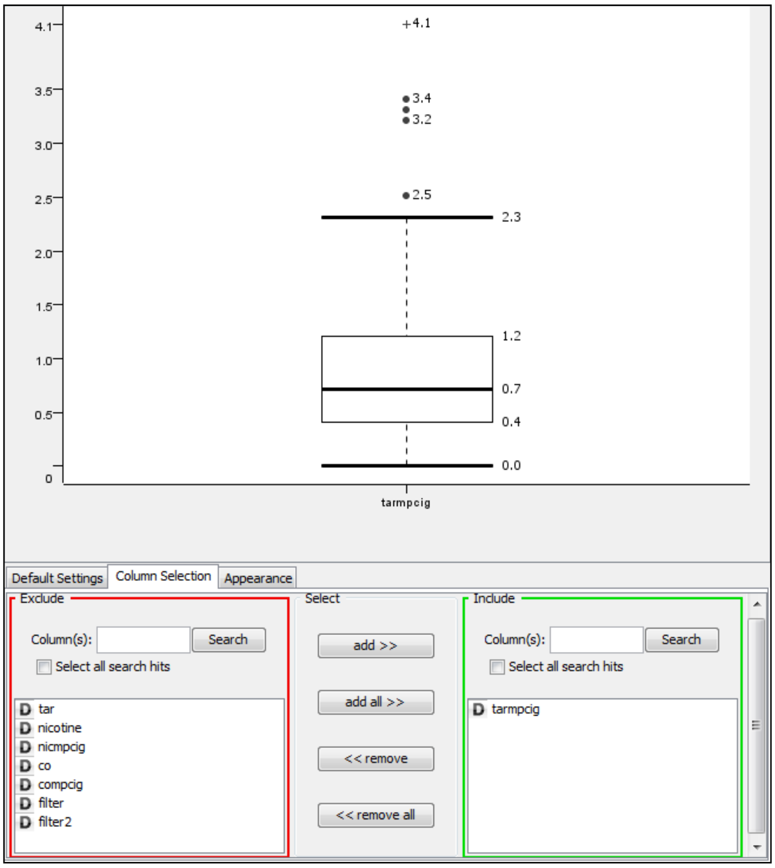


Figure 13-23: A boxplot.

- **Conditional boxplot:** Boxplots for several groups (such as geographic regions) can be placed side by side on a single graph for easy comparison, as shown in Figure 13-24.
- **Parallel coordinates:** The plots show values for several variables all together on a single plot, with the values for each case connected by line segments. Common combinations stand out from the rest. For example, look at Figure 13-25, which shows several variables related to cars and fuel consumption. Many cases share certain values, exactly or approximately,

forming dark patterns from the many lines following similar paths across the graph. For example, cases for cars with four cylinders, low displacement, high mileage, and late model years form a very dark and conspicuous pattern. (Parallel coordinates plots are hard to read when too many cases are used. If this happens to you, take a random sample of your data, as shown in Figure 13-26, and make a new plot.)

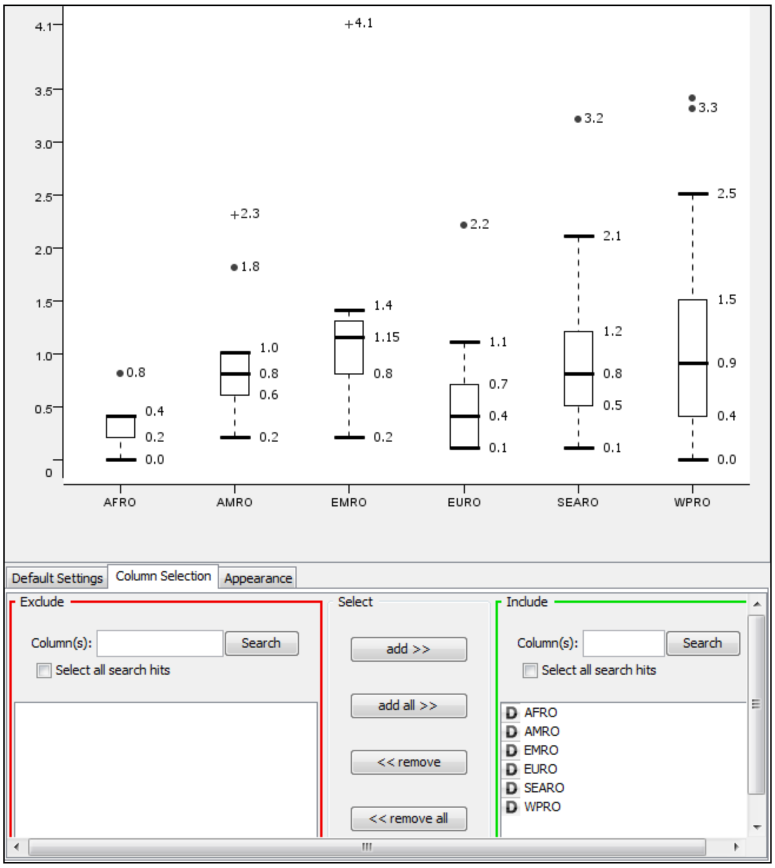


Figure 13-24: A conditional boxplot.

- **Gains charts (also called cumulative gains):** A gains chart (see Figure 13-27) shows you how much a predictive model improves results over random sampling. Some people are more likely to take action (buy a product, vote for a candidate, break the law . . .) than others. If you know nothing about a group of people, the best you can say is that contacting half of the people will turn up half of those who will take action. But a predictive model can tell you which people are the best prospects, so you can use the model to pick half (or 10 percent or 60 percent, and so on . . .) and get more action. How much more? In the chart in

Figure 13-27, you can see a diagonal line where the *x* and *y* values are always the same; this represents what you'd get by selecting prospects at random. The other line represents the model. The difference in *y* values between the model and the random selection shows how much the model improves your outcome. Read the model line plotted on the chart, and compare it to the line for random sampling.

- **Lift charts:** Lift charts are very similar to gains charts. The key difference is that the data is normalized, so that random sampling is always represented as a value of 1 and model results are shown in proportion to random sampling. (Refer to Figure 2-29 in Chapter 2 for an example.)

You may see several different types of charts called lift charts. Some are cumulative, and others are not. Some may even be gains charts (described previously).

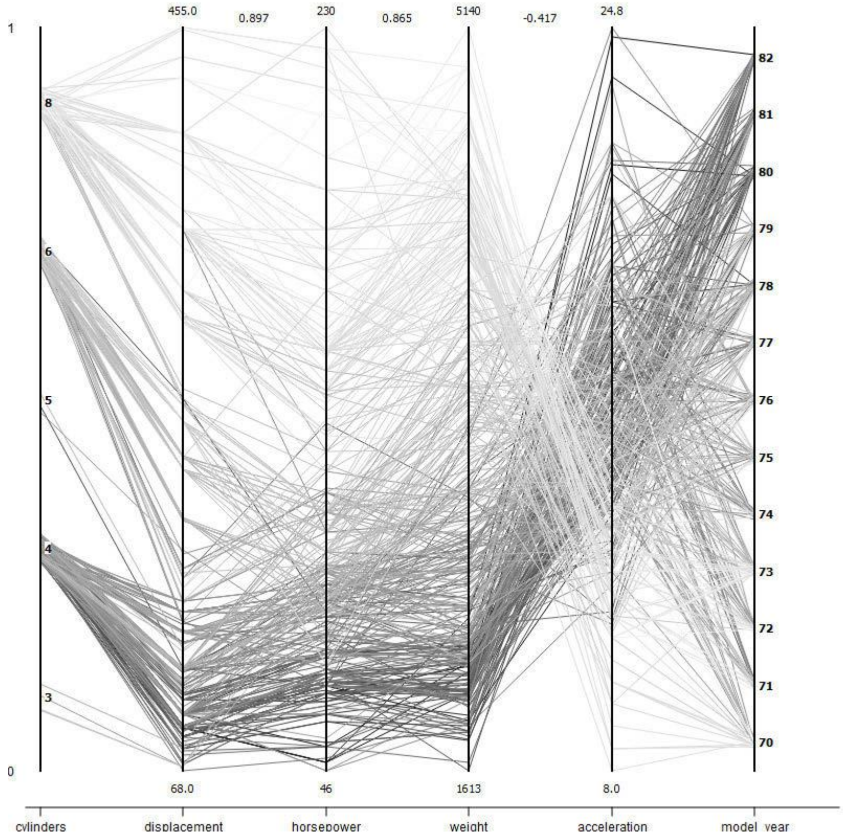


Figure 13-25: A parallel coordinates plot.

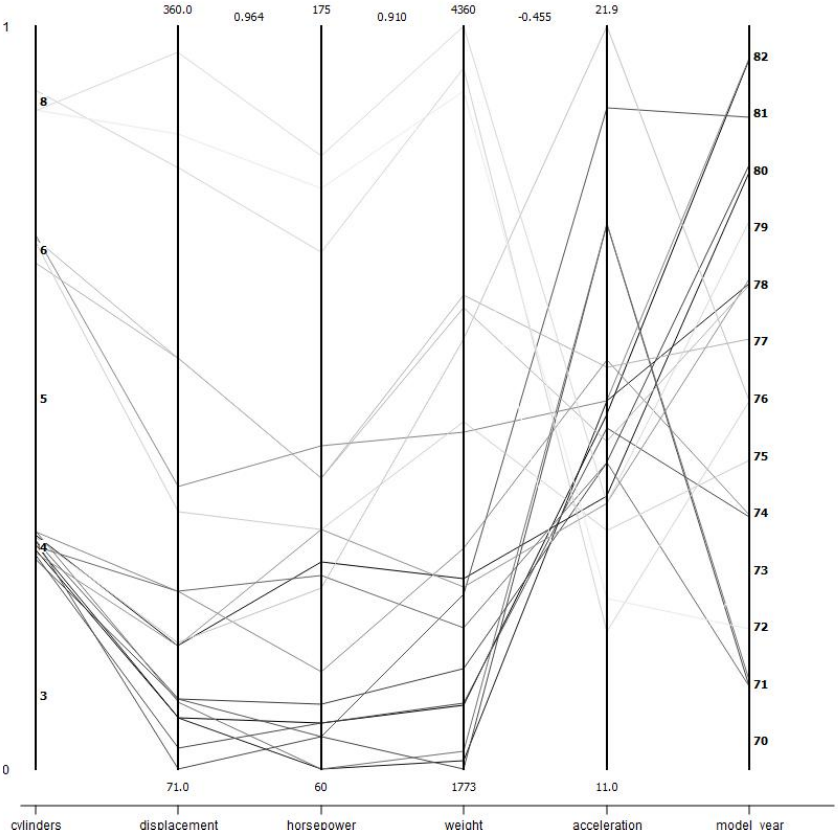


Figure 13-26: Sampling to reduce the number of cases for a plot.

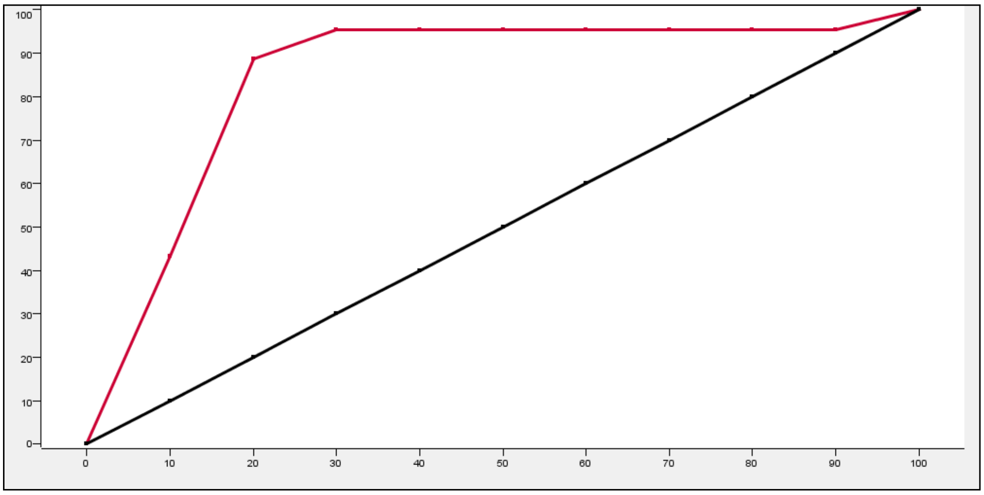


Figure 13-27: A gains chart.

Putting graphics in context: An interview with Laura Kippen

Graphs are outstanding tools for communication, and some are lovely to look at, but don't assume that others will immediately see what you see in them.

Laura Kippen integrates graphs with commentary and tables in her reports and live presentations. She's the founder and president of InfoManiacs, a full-service qualitative and quantitative marketing research firm. Laura helps clients make business decisions by using advanced analytics to provide relevant, actionable information.

Q: How important is the final report or live dialogue in the data-mining process?

Kippen: Presentation of information, especially information that was derived utilizing some form of advanced analytics, is critical to the success of a project. Let's face it, if your clients don't get it, they aren't going to be happy.

Q: What role do graphics play in presentation?

Kippen: I don't recall a moment when the chart, graph, or table itself made the light bulb go on. Typically what I have experienced is that a chart or graph elicited a response, but it had more to do with how I highlighted the information and then explained why it was important and not the chart or graph itself.

Q: So, can data miners do without graphs?

Kippen: It's not really my charts that make it all happen; it's all the components of a report that are carefully considered that help clients see the entire story. However, that said, I would never be able to develop and make a compelling story without well-considered graphs, charts, and tables.

Q: Do you find any kinds of graphs that are particularly powerful?

Kippen: Flashy graphics may dazzle but not add information or say anything new. In other words, you can usually make the same point using standard, pedestrian graphs and tables, and clients are likely to respond favorably. At the end of the day, if your report doesn't say anything of value, all the flashy graphics in the world won't save you.

Q: Many data miners feel that they aren't giving clients their money's worth if they don't use fancy graphics. Clearly, you don't buy that. How does your resistance to using glitzy graphics go over with clients?

Kippen: My clients like what I give them and they come back for more.

So, here are the lessons that data miners can learn from Laura:

- Conventional graphs can be effective when used well.
- Eye-catching graphics don't necessarily add new information.
- Graphs may be only part of what makes an informative report for a client — but they're an indispensable part.