An Introduction to Visual Multivariate Analysis

Stephen Few
July 11, 2006

The analysis of business data can be classified into several types. These types of analysis vary based on the nature of the data and the specific relationships that you attempt to discover and understand. For example, one of the most common types is *time-series analysis*. In this case, the information spans some range of time that has been divided into intervals, such as years, months, days, or hours, and all of the relationships that interest you involve patterns of change. Each type of analysis focuses on particular patterns that represent meaningful characteristics of the data. These patterns can be presented in visual form by giving the data shape, as well as other attributes (for example, color) that can be seen, examined, and understood. Time-series analysis strives to understand patterns such as trends, rates of increase or decrease, volatility or stability, leading and lagging indicators, and seasonality.

Typical business data analysis rarely ventures beyond the investigation of time-series, ranking, part-to-whole, and deviation relationships. Many rich finds live outside this familiar territory, however, and are worth the journey. One that receives a great deal of attention in the scientific community, but only superficial attention in the business community except among small cadres of statisticians, is *multivariate analysis*.

Our data warehouses and other repositories, including spreadsheets, store many attributes of the things that concern us, such as products, customers, and orders. Using products as an example, we characterize them based on several attributes, such as product type, price, age (such as the number of months or years on the market), units sold, revenue, production costs, marketing costs, profit, and customer satisfaction. Another name for attributes like these is *variables*—characteristics that can have various values. When we wish to understand the relationship between some set of variables that characterize a product (or customer or whatever) or how the profile formed the values of these variables for a particular product compares to the multivariate profiles of other products, we call this process of investigation *multivariate analysis*.

The traditional business intelligence tool for multivariate analysis (or *multidimensional analysis*) is the pivot table, also known as the crosstab. These interactive text-based tables of data have helped us uncover marvelous treasures, but visual tools can extend our ability to think about the data multidimensionally. Some important relationships that live in our data only assert themselves when given shape before our eyes.

Multivariate analysis is especially useful for three lines of investigation. Using the product example mentioned earlier, these three avenues of pursuit can be expressed as the following three questions:

- What correlations exist between the variables that characterize our products?
- Which of our products have similar multivariate profiles and can therefore be grouped together?
- What conditions in other variables most contribute to a particular effect in this variable that is being examined (for example, high profits)?

Each of these three lines of investigation can be most effectively pursued using particular types of visualizations and particular techniques for interacting with multivariate data. In this article, I will list and briefly introduce a few of the visualizations that bring to light the relationships between sets of variables. In subsequent articles later in the year, I'll explore several of these multivariate visualizations in detail, one at a time.

The following is a list of the multivariate visualizations that I've encountered in my work that have merit:

- Crosstab arrangements of small multiples (a consistent series of small graphs)
- Multiple concurrent views with *brushing* functionality
- Heatmap matrix
- Parallel coordinates
- Glyphs
- The table lens

Crosstab Arrangements of Small Multiples

Edward Tufte introduced the term *small multiples* many years ago to describe an arrangement of small graphs, all within eye span, which look precisely the same (including a consistent quantitative scale), except that each displays a different subset of a larger set of data. These subsets of data vary along a selected variable, such as regions, with one graph per region. For example, imagine a series of bar graphs that each displays product revenue for an entire year (one bar per quarter), but each does so for a different sales region. This arrangement of small multiples allows us to add another variable (dimension) to the display, making it more multivariate by one, without harming the clarity of the picture, which would not have been achieved by using a single graph and making it three dimensional.

This display can be made even more multivariate by arranging the small multiples as a crosstab. In Figure 1, you see a set of small multiples arranged as a crosstab, which extends the number of variables by one more. In addition to the sales regions, which vary per column, a separate product is now displayed in each row. Actually, I included another variable along with product by assigning colors to the bars to encode profits as well, ranging from negative values in red (darker indicates more negative) to positive values in green (darker indicates more positive). Extremes on both ends of profits (Colombian coffee in the East and Green Tea in the West) stand out clearly.

Figure 1: Example of a crosstab arrangement of small multiples, created with Tableau Software.

Multiple Concurrent Views with *Brushing* Functionality

Another approach to multivariate analysis involves the ability to place several different views of the same data set (multiple graphs, tables, etc.) on a single screen, complemented by a technique called *brushing*, which allows you to select a subset of data in one of the displays (for instance, by clicking on a bar in a bar graph) to highlight it, resulting in that same subset of data being automatically highlighted where it appears in every one of the views. This provides a powerful means to examine connections between multiple variables.

Figure 2 provides a simple example of this approach, showing three views of revenue data: one by region, one by industry, and another that correlates margin and revenue in the form of a scatterplot. By brushing the low margin sales in the scatterplot, I can see that they fall to an excessive degree in the healthcare industry, yet their regional distribution seems to be fairly consistent with overall sales per region.

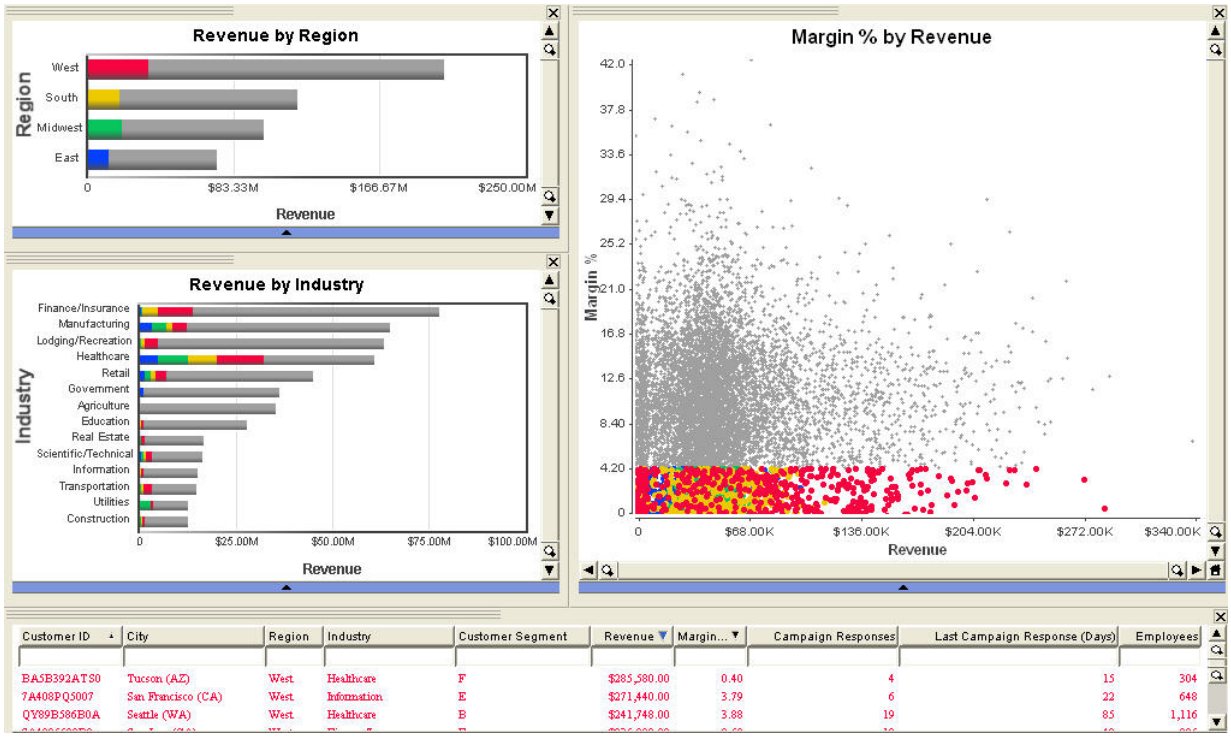

Figure 2: Example of a display, created with Advizor Analyst, which includes multiple views of the same data set, complemented by the highlighting technique called *brushing*.

Heatmap Matrix

The term *heatmap* generically refers to any visual display that uses variations in color to encode a quantitative variable. We are all familiar with weather maps, which use color to encode variations in temperature or precipitation. A heatmap matrix, then, is a tabular arrangement of cells that each encodes a quantitative value as color corresponding to some categorical (non-quantitative) variable across the columns and another categorical variable down the rows. In Figure 3, you see sales data, with sales regions and states along the columns, products down the rows, and revenue as color ranging from low revenues (light gray) to high revenues (dark gray).

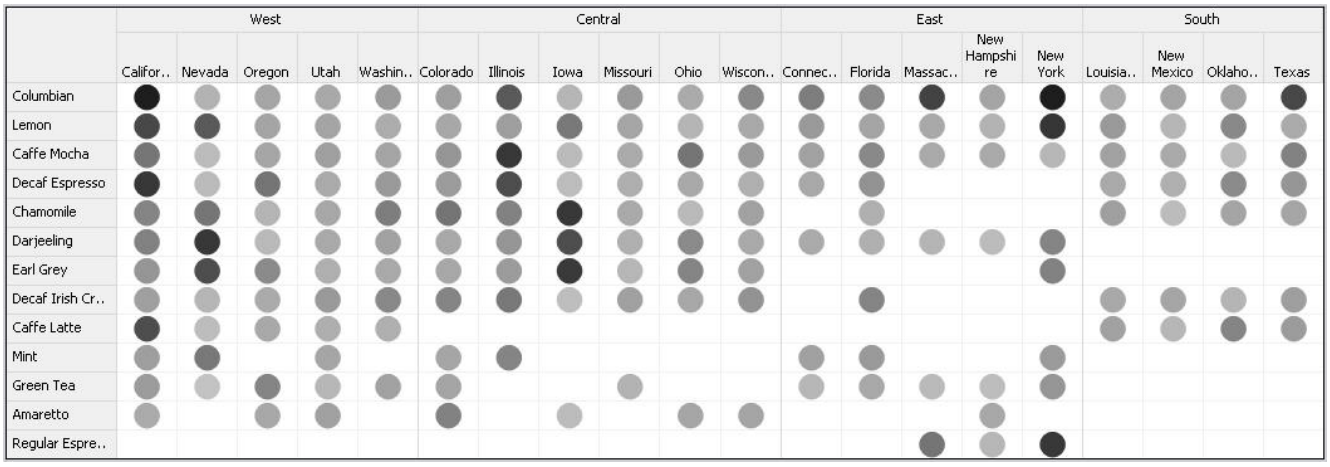

Figure 3: Example of a simple heatmap, arranged as a crosstab, created with Tableau Software.

We can add more variables to a heatmap matrix by deviating from the strict crosstab arrangement and placing an entirely different variable in each column. In Figure 4, you now see the same row-wise series of products, but each column now represents a different variable. Each column uses a range of color to encode its range of quantitative values, but the scale of each differs based on its unique set of values.

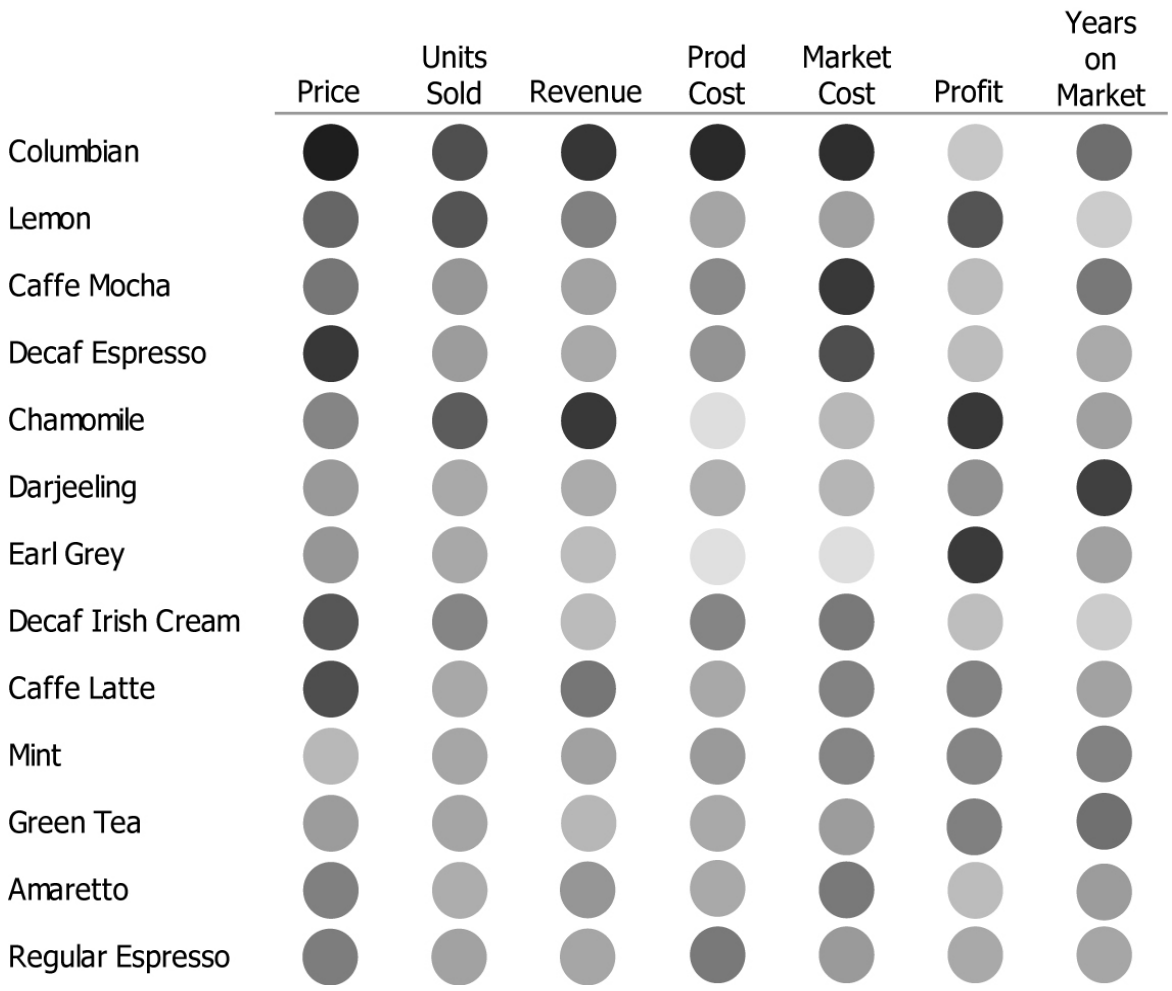

Figure 4: Example of a multivariate heatmap matrix.

You might find yourself objecting to the fact that color doesn't encode quantitative values in a way that allows precise decoding and comparison. You are absolutely right. Even when designed properly, color can only provide a rough approximation of value and magnitude comparisons. When analyzing large data sets (usually much larger than what's shown in this example) across multiple variables, we view it from the 30,000-foot level, hoping to spot predominant patterns and exceptions. At this level, we don't make precise comparisons, which would require different visualizations and much less data.

Parallel Coordinates

This is a visualization that is probably new to most of you. It tends to be used primarily by scientists. The first time people see parallel coordinates, they usually cringe at what appears to be an absurd and meaningless clutter of lines. Take a look at Figure 5 and see what you think.

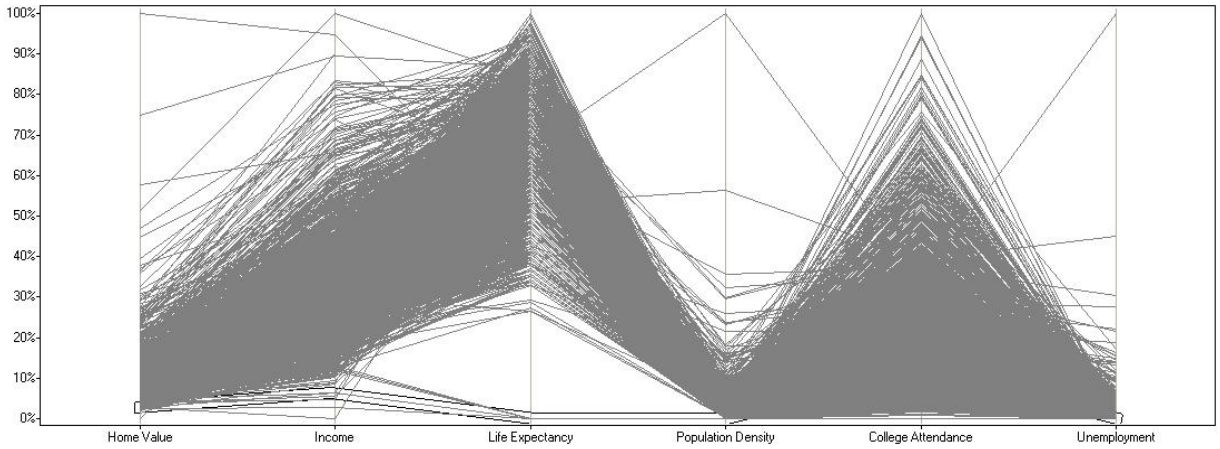

Figure 5: Example of a parallel coordinates display, created with Spotfire DecisionSite.

Now, if you've finished laughing, let me explain what you're seeing in an attempt to bring some order to this chaos of lines. Each line, running from left to right across the entire plot area is a separate entity, in this case a U.S. county. What you're seeing is a line for every county in the entire United States of America, totaling 3,138. Each vertical axis, spaced evenly across the plot area from left to right, represents a different variable (home value, etc.), with its full range of values sequenced along the axis from bottom to top. The line for any given county encodes several values that characterize that county where it intersects each axis. With a line for every county, the display is cluttered, to say the least. But even from this clutter, when viewed from our 30,000-foot perspective, particular patterns and exceptions emerge. For example, most counties have home values that are 30% or less compared to the county with the highest home values, and the population density in one county is almost twice that of any other county.

The real power of parallel coordinates comes from interactions with the display to filter out what doesn't interest you and to find all of those entities that match a particular multivariate profile that does interest you. In Figure 6, I have selected only those counties with the highest levels of college attendance and filtered out all other counties, which makes it easy to see that these counties correlate strongly with low rates of unemployment and low population densities.

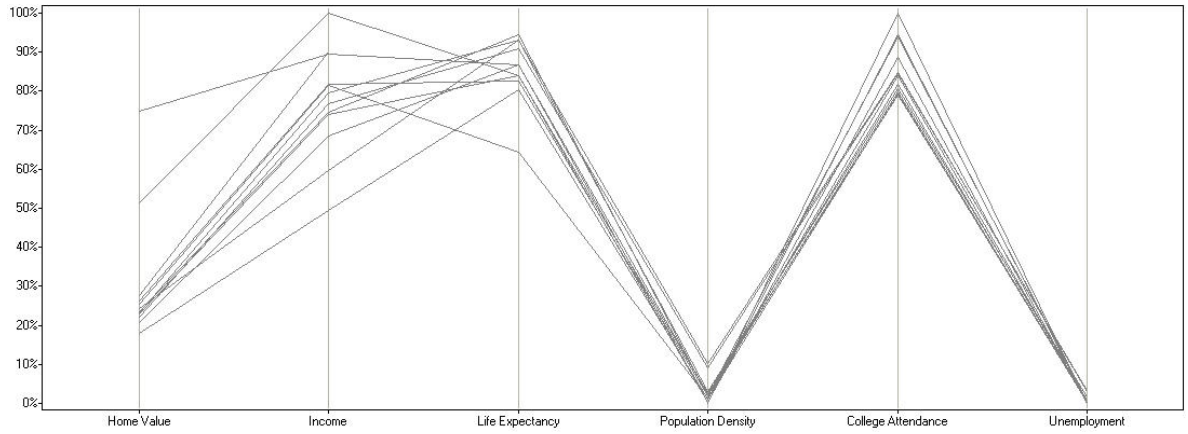

Figure 6: Example of a parallel coordinates display, created with Spotfire DecisionSite, which illustrates what can be discovered when data is filtered out to focus on particular features.

Glyphs

The term *glyph* refers to a graphical object that simultaneously represents the values of multiple variables. Glyphs can be quite complex (sometimes absurdly complex) and can exceed the bounds of meaningful display; but when properly designed, applied to the right multivariate data, and used for an appropriate purpose, glyphs can paint a useful picture, especially for monitoring a large set of data for irregularities or specific patterns of interest.

One of the simplest examples is called a *star glyph*, because its shape resembles a star. To construct one, begin with a radial arrangement of axes that extend from a common center point outward to form a circle, much like a radar chart. Each axis represents a separate variable with low values near the center and high values near the perimeter. Imagine that each glyph represents several measures that characterize a rider in the Tour de France. The axes represent pulse rate, body temperature, blood pressure, respiration, amount of water consumed, miles ridden, miles per hour, and cycles of the bicycle's wheel per minute. Now, connect the dots to form an outline of the various values and then erase the axes, so that only the shape formed by the outline remains. Your result will be similar to the glyph on the right in Figure 7. This little glyph represents the condition and performance of a single rider. Finally, fill your computer screen with enough of these small glyphs to represent every rider in the race.

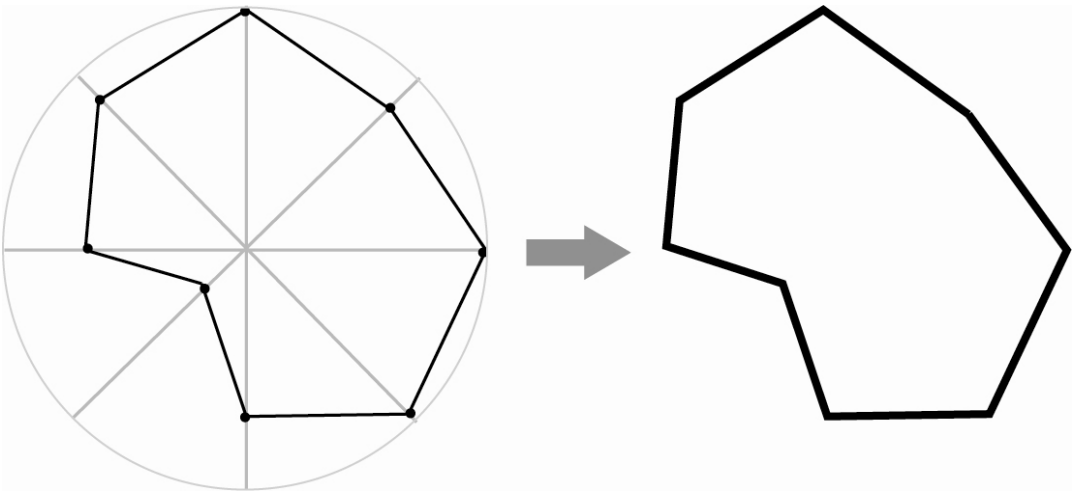

Figure 7: Example of a star glyph on the right, along with a version that displays how it is constructed using the axes on the left.

Once you've become intimately familiar with this particular glyph and know which shapes represent particular conditions of concern, you might be able to monitor a screen full of glyphs and notice when particular riders are exhibiting problems. You could respond by dispatching a medical team to be at the ready should matters worsen. After the race, you could isolate the top 10 riders and replay only their glyphs together, observing how they changed through time to see if these riders exhibited any similar conditions that might have contributed to their successful performance. You could do the same for those riders who performed poorly in an effort to identify problems.

One of the keys to effective glyph displays is to keep them simple, encoding the data in ways that highlight the most meaningful conditions in ways that can be seen and understood at a glance. I've seen glyph displays fizzle to a well-deserved demise because they failed to observe this perceptual rule.

The Table Lens

I'll say little about this final visualization, not because it deserves little attention, but because Ramana Rao, a founder of Inxight Software, Inc., has written an entire article to describe this form of multivariate display. Ramana Rao, along with Stuart Card, developed the table lens when they worked together at Xerox PARC (Palo Alto Research Center). Now that you have an understanding of multivariate analysis and the visualizations that support it, read Ramana's article, *TableLens: A Clear Window for Viewing Multivariate Data,* for a full introduction to the table lens, which he knows so well.

About the Author

Stephen Few has worked for over 20 years as an IT innovator, consultant, and teacher. Today, as Principal of the consultancy Perceptual Edge, Stephen focuses on data visualization for analyzing and communicating quantitative business information. He provides training and consulting services, writes the monthly *Visual Business Intelligence Newsletter*, speaks frequently at conferences, and teaches in the MBA program at the University of California, Berkeley. He is the author of two books: *Show Me the Numbers: Designing Tables and Graphs to Enlighten* and *Information Dashboard Design: The Effective Visual Communication of Data*. You can learn more about Stephen's work and access an entire library of articles at www.perceptualedge.com. Between articles, you can read Stephen's thoughts on the industry in his blog.