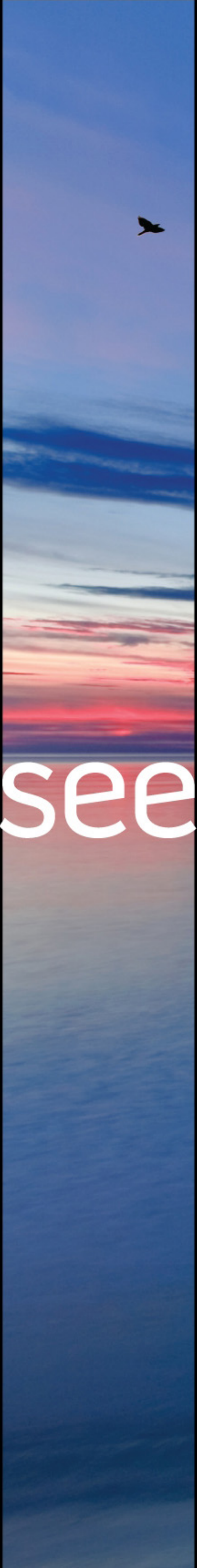
Simple Visualization Techniques
for Quantitative Analysis

Now you see it



STEPHEN FEW

12 MULTIVARIATE ANALYSIS

Introduction

Multivariate analysis is a bit different from the other types we've examined. As with all types of quantitative data analysis, the fundamental activity is comparison, but the comparison in this case is more complex than for other types. Other forms of analysis usually compare multiple instances of a single quantitative variable, such as sales revenues per region, or one variable to another, such as revenue to profit. In contrast, multivariate analysis compares multiple instances of several variables at once. The purpose of multivariate analysis is to identify similarities and differences among items, each characterized by a common set of variables.

A simple example involves automobiles. Imagine that we work for an automaker, and we're trying to determine which characteristics contribute most to customer satisfaction for particular types of buyers. Our database includes the following variables, which we'll use to characterize and compare each of the cars:

- Price
- Gas mileage
- Top speed
- Number of passengers
- Cargo capacity
- Cost of insurance
- Repair costs
- Customer satisfaction rating

The values of all these variables for each car combine to form its multivariate profile. We want to compare the profiles of cars to find which ones best characterize each type of buyer, from those looking for a basic commuter vehicle to those looking for thrills. To do this effectively, we need a way to see how the cars compare across all selected variables at once. We must compare these variables as whole sets, not just individually. Multivariate analysis revolves around the following questions:

- Which items are most alike?
- Which items are most exceptional?
- How can these items be combined into logical groups based on similarity?
- What multivariate profile corresponds best to a particular outcome?

The patterns that answer these questions are the ones that are most meaningful in multivariate analysis.

Multivariate Patterns

In multivariate analysis, we examine patterns formed by several values that measure different attributes of something, which exhibit its *multivariate profile*. To do this, we must find ways to represent several variables worth of information about something as a single composite pattern. To pursue the questions listed above ("Which items are most alike?", and so on), multivariate profiles must be displayed in a way that makes it easy for us to spot similarities and differences among them even when hundreds of items are in play. The form that these patterns take is determined by the type of visualization that we use, which is what we'll take a look at now.

Multivariate Displays

I'll introduce three quite different displays that people have attempted to use for multivariate analysis. Only one is truly effective; of the other two, one does the job poorly (I've included it only so you know to avoid it) and one works satisfactorily when a better means isn't available.

These three types of displays go by the following names:

- Glyphs
- Multivariate heatmaps
- Parallel coordinates plots

The method that works well is the parallel coordinates plot, but be forewarned: it will probably seem absurd and overwhelmingly complex at first glance.

Glyphs

You can guess, based on the fact that it's part of the word "hieroglyphics," that a glyph is a picture of something. Egyptian hieroglyphics were pictures that formed a written language. In the context of information visualization, the term "glyph" has a particular meaning: "A glyph is a graphical object designed to convey multiple data values."[1] A glyph is composed of several visual attributes, each of which encodes the value of a particular variable that measures some aspect of an item. To illustrate how glyphs work, I'll construct one from scratch that I doubt has ever been used for multivariate analysis (and I hope will never be). I'll use stick drawings of people to represent multiple variables that describe or measure aspects of human physiology and health. Each of the following three glyphs represents the health of a different individual:

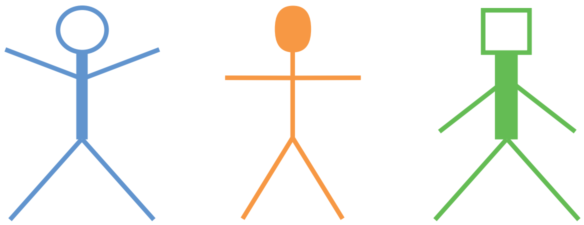


Figure 12.1

[1] *Information Visualization: Perception for Design*, Second Edition, Colin Ware, Morgan Kaufmann Publishers, San Francisco CA, 2004, p. 145.

The variables are encoded as follows:

Visual Attribute	Variable
Color	Body temperature
Shape of the head	Blood type
Thickness of the torso	Body mass index
Position of the arms	Heart rate
Position of the legs	Blood sugar level

If we memorized the meanings of each visual attribute and learned how to decode particular expressions, such as a square versus a round head or a wide versus a narrow stance, we could theoretically use these glyphs to examine and compare the health of many individuals.

The best known example of a glyph was created by Herman Chernoff in 1972. He used simplified line drawings of the human face and mapped different variables to particular facial features (size of the eyes, curvature of the mouth, shape of the head, and so on). Here's a sample collection, each a little different.

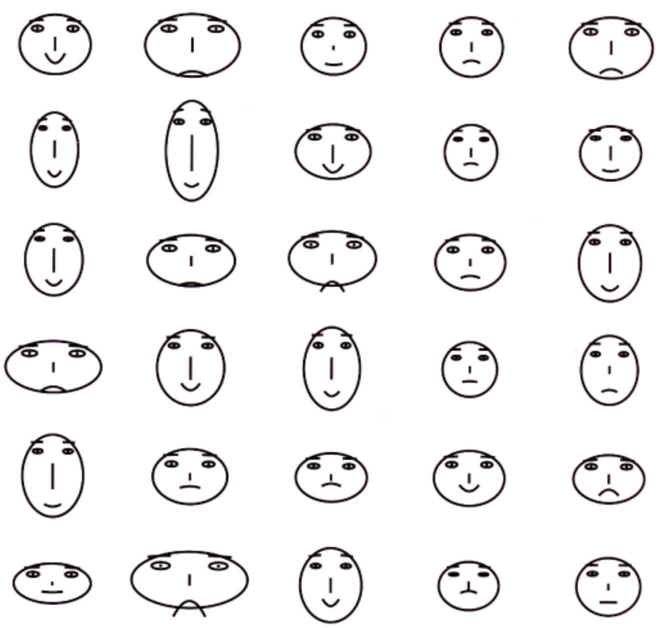


Figure 12.2. Chernoff introduced this idea in an article, "Using faces to represent points in *k*-dimensional space," *Journal of the American Statistical Association,* 68, 1973, pp. 361-368.

Chernoff chose the human face because human perception has evolved to rapidly read and interpret facial expressions. From early childhood we learn to recognize faces and respond to subtle facial expressions although much more is communicated by facial expression than we usually learn to recognize, such as whether or not someone is telling the truth. Despite a great many research studies that have used Chernoff faces, I have never seen any convincing evidence that they work effectively for multivariate analysis.

Two other glyphs that have also been used for multivariate analysis are called *whiskers* and *stars*. Whisker glyphs, illustrated below, consist of multiple lines that radiate out from a center point. Each line represents a different variable and its length encodes its value.

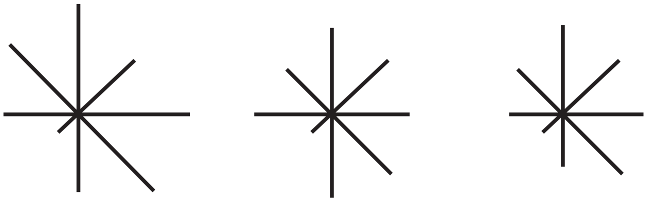


Figure 12.3

Star glyphs, illustrated below, are similar to whiskers in that variables are encoded as distance from a center point. This time the endpoints of the radiating lines are connected to form an enclosed shape.

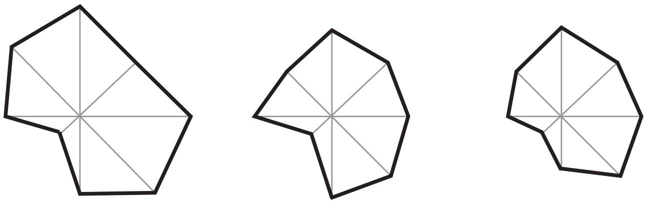


Figure 12.4

I mentioned before that we could theoretically use glyphs to examine and compare multivariate profiles. But what works in theory does not always pan out in practice, and I believe that this is one of those cases. If you ever encounter a product that promotes glyphs for multivariate analysis, don't let the novelty of the display entice you. Put it to the test, attempting to use real data to solve real problems, and see if it works. Until glyphs can prove their worth in the realm of real-world data analysis, I recommend that you avoid products that are based on this approach.

Multivariate Heatmaps

In general, heatmaps are visual displays that encode quantitative values as variations in color. Sometimes when we speak of heatmaps, we're referring to a matrix of columns and rows, similar to a spreadsheet, that encodes quantitative values as color rather than text. Heatmaps, such as the following example, can be used to display multivariate data. In this case the items on display are products (one per row) and the quantitative variables (one per column) consist of the following:

- Price
- Duration (length of time on the market)
- Units Sold
- Revenue
- Marketing Expenses
- Profit

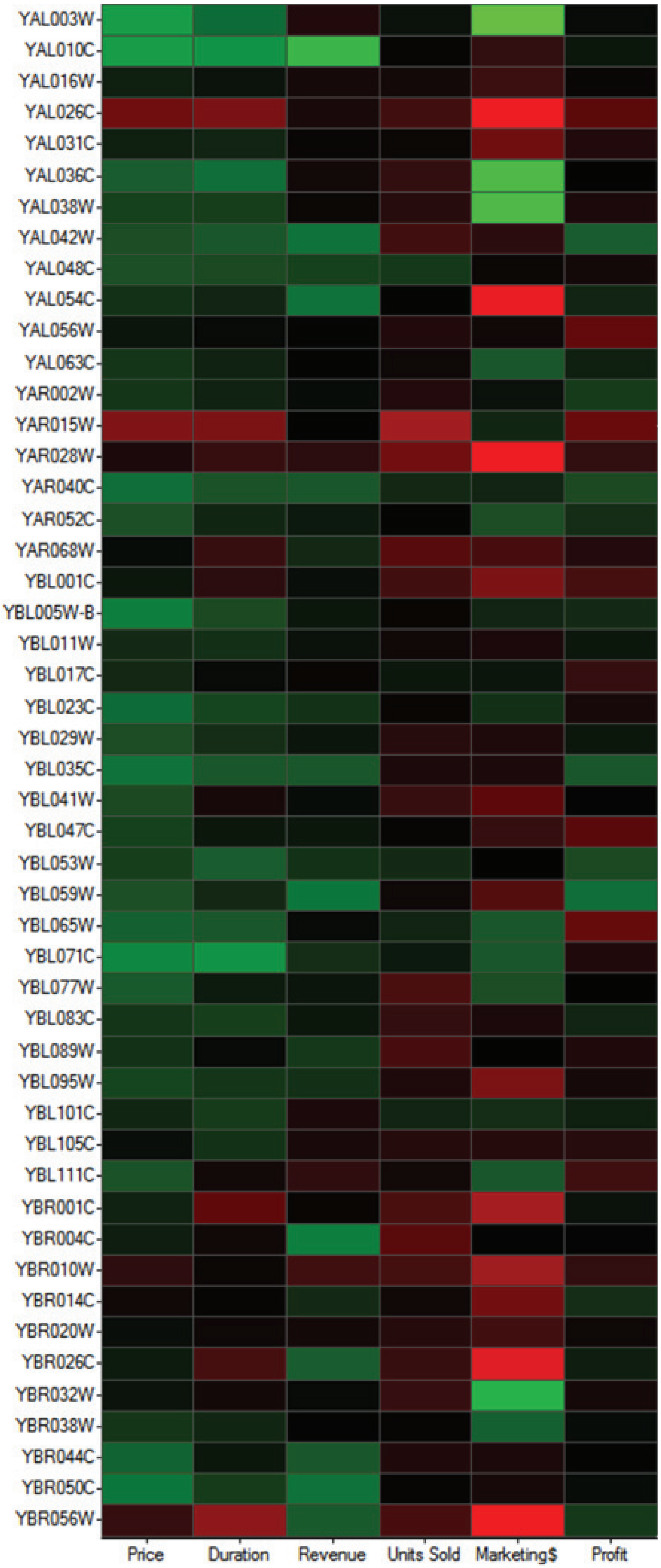


Figure 12.5. Created using Spotfire

The combination of colors across a single row displays a product's multivariate profile. In the previous example, higher-than-average values appear as green (the darker the higher), near-average values appear as black, and lower-than-average values appear as red (the darker the lower).

As you can see, heatmaps alone can be difficult to use when searching for similar profiles, but they can be used to reveal exceptions, such as the bright green and bright red values in the Marketing$ column, and predominant multivariate patterns, such as the fact that product YAL026C (the fourth row) exhibits all lower-than-normal values (red), except for Revenue.

Whether they're being used to display multivariate data or for other purposes, heatmaps suffer when colors haven't been chosen wisely. The previous example illustrates two common problems:

- The distinction between red and green cannot be seen by the 10% of males and 1% of females who suffer from the most common form of color blindness.
- When multiple hues are appropriate for encoding continuous values, such as positive and negative numbers, a dark color such as black usually shouldn't be used to encode values in the middle (in this case values near average), because it is much too salient (that is, visually prominent). Assuming it's appropriate to encode these values as above and below a particular value (in this case as above and below the average value for all products), the colors in this next version work better. Positive values are blue, negative values are red, and values near zero (average) fade from blue and red to light gray. No form of color blindness that I know would prevent people from seeing the difference between red and blue. The light gray that has been used to represent numbers close to zero intuitively represents low values and grabs our attention less than the vibrant reds and blues that have been used to draw our attention to extremes on both ends of the continuum.

Better colors have improved the following heatmap, but its usefulness for multivariate analysis is still limited, because it's difficult to see the combination of colors for particular products as a pattern. Because multivariate profiles are complex by nature, no visualization can display them with perfect clarity or be analyzed with utter ease, but the one we'll examine next stands a full head above the others.

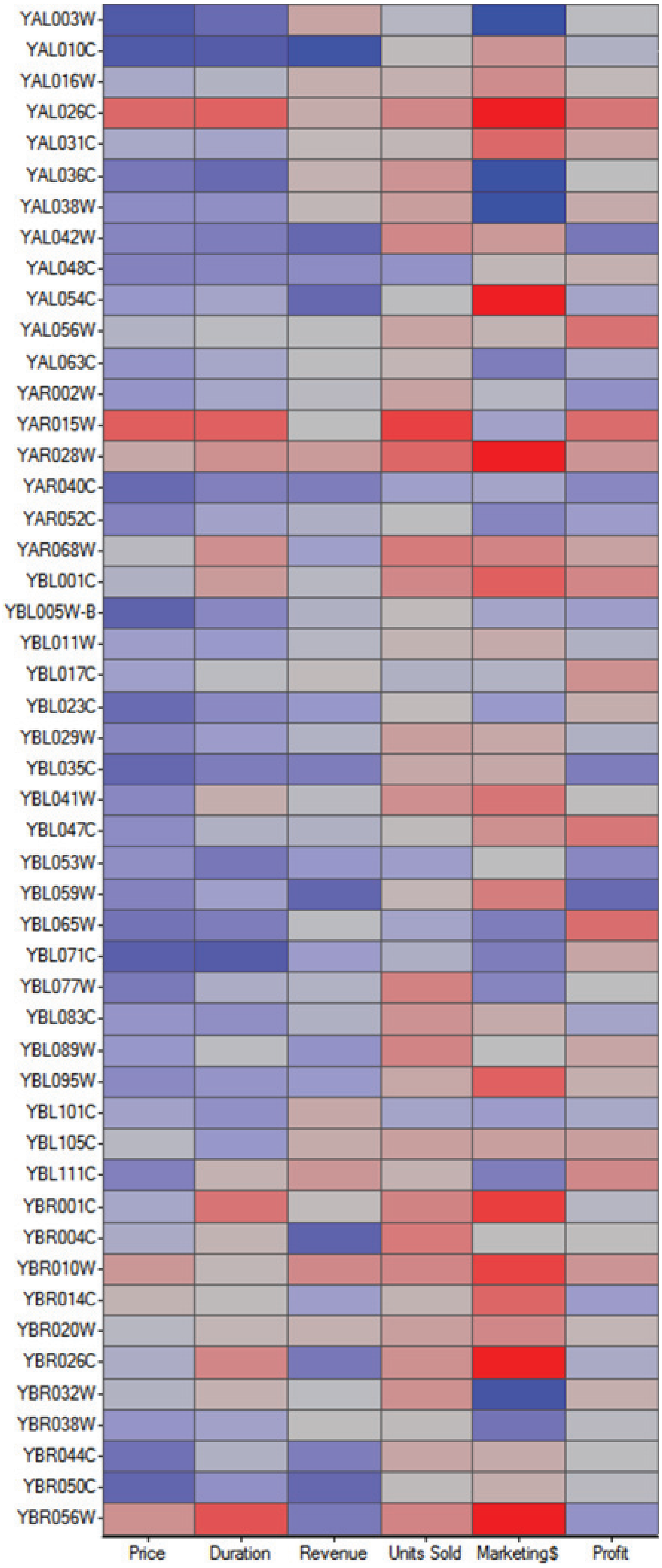


Figure 12.6. Created using Spotfire

Parallel Coordinates Plots

The first time I laid eyes on a parallel coordinates plot, I laughed and cringed simultaneously because it struck me as a ridiculously complex and ineffective display. Ordinarily, if graphs that use lines to encode data include more than a few, they deteriorate into useless clutter. Parallel coordinates plots, however, can include hundreds of lines, which in most cases would boggle the senses. Even the example below, which includes only 49 lines, will likely strike you as absurd.

Parallel Coordinate plots were invented by Alfred Inselberg in the late 1970s.

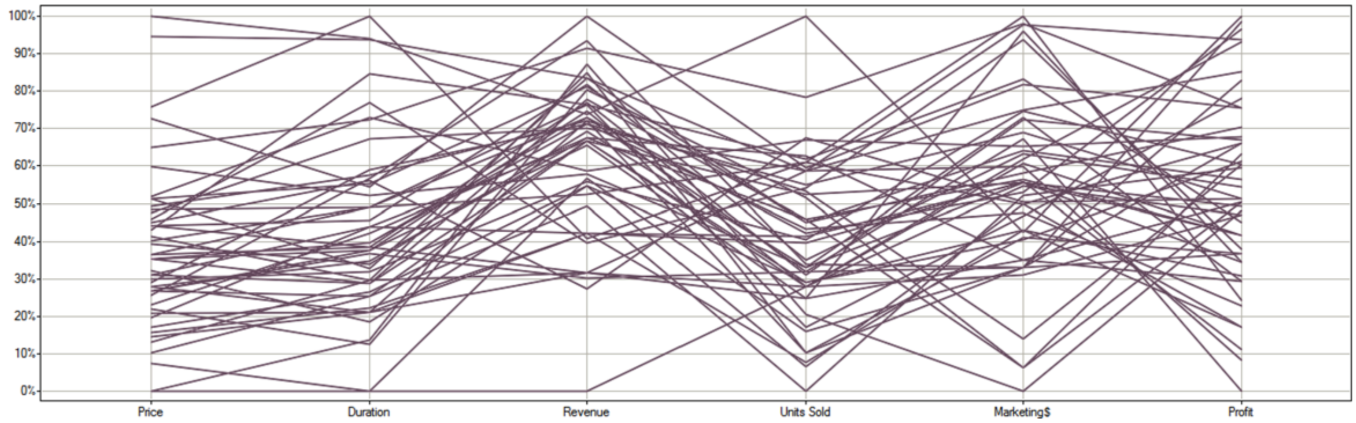


Figure 12.7. Created using Spotfire

However, this frenzy of intersecting lines, when used properly for multivariate analysis, can actually lead to great insight. It includes six variables (price, duration on the market, revenue, units sold, marketing expenses, and profit) for 49 products, one product represented by each of the lines that extends from left to right across the graph. The values for each variable in this particular example are laid out along percentage scales, which all begin at 0% at the bottom and extend to 100% at the top. For example, product prices, rather than being expressed in dollars, have been converted to percentages, with the highest priced product equal to 100% and the lowest equal to 0%. Parallel coordinates plots can be set up this way, using a common scale for all variables, or so that each variable is scaled independently using the original values, such as dollars, counts, and so on. Either way, the display looks and works in basically the same way.

Unlike regular line graphs, which should only be used to connect values along an interval scale such as time, parallel coordinates plots connect values associated with entirely different variables. In this example, measures of six different variables for a single product are connected with a line that intersects each axis at the point where the product's value for that variable is located along the scale. Although it's unusual to display values across multiple variables using a line to connect them, the shape formed by the line is meaningful because it forms a multivariate pattern that describes a particular product. The individual patterns formed by different lines can be compared to determine similarities or differences among products. That is, they could be compared if it were possible to distinguish one line from another, which we can't do with the display above.

Unlike line graphs, which use lines to connect values along an interval scale such as time, patterns formed by individual lines in a parallel coordinates plot contain less inherent meaning. If we change the order of the variables in the plot, the patterns will change, so we usually shouldn't become too attached to particular patterns, such as by taking time to memorize them as significant, but should use them only to examine and compare multivariate profiles in the moment.

What we can do with this parallel coordinates plot in its current state is glimpse the big picture: predominant patterns and exceptions. For example, we can see that one of the products (YAL010C), represented by the line that I've highlighted below, has a revenue amount that is much lower than the others. We can speculate that this is due to its short lifespan (notice its duration value) and perhaps, in part, to its low price.

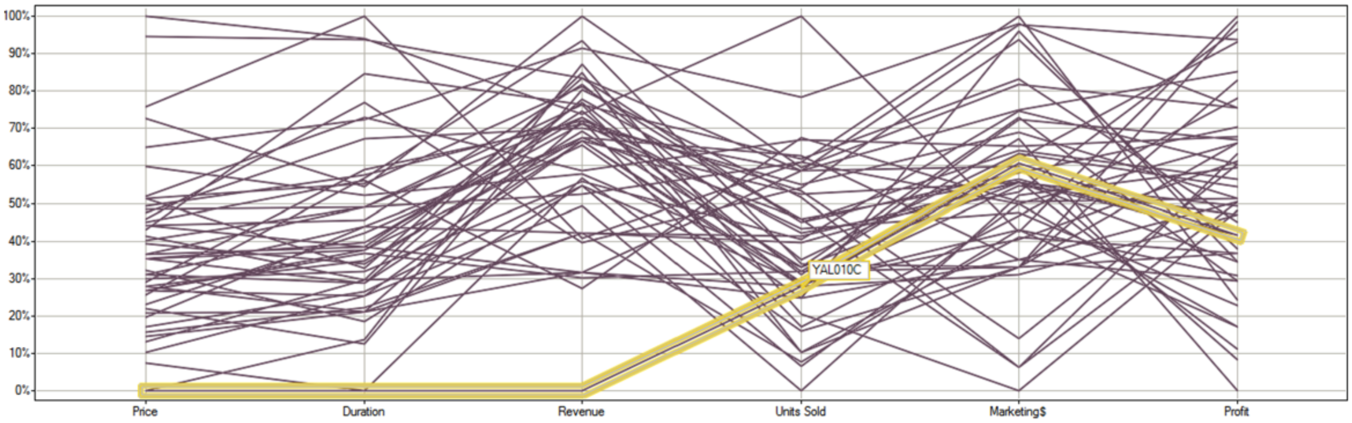


Figure 12.8. Created using Spotfire

We can make a few other observations at this point:

- One product has sold quite a bit more than its closest rival.
- Most products have been on the market from between 20% and 60% of the full range of time for all products.
- There is a heavy concentration of revenues between 65% and 85% of the full range, from least to greatest.

Even in the midst of clutter, predominant patterns and exceptions are visible. More detailed understanding will, however, require interaction with the data to cut through the clutter.

Most often when analyzing multivariate data, we look for multivariate profiles that correspond to a particular condition. For instance, if we were examining the current example, we might want to find out which multivariate profiles most correspond to high profits. It helps to position the primary variable of interest, such as profits in this case, as the last axis, as I've done in the example above, because this makes it easier to focus on that variable. To investigate multivariate conditions associated with high profits thoroughly, however, we'll need to focus on the products that intersect the profit scale near the high end, for example those with profits of 80% and above. Although we can easily distinguish those lines where they meet the profit axis, it is still difficult to follow them across the other axes because all products are displayed as dark gray lines, which are hard to distinguish from one another. So, for our analysis, we need some way to clearly separate products with high profits from the others. This can be accomplished using brushing and filtering, two of the methods that we examined in *Chapter 4: Analytical Interactions and Navigation*.

In the following example, all products with profits above 80% have been brushed, changing their color to red to make them stand out:

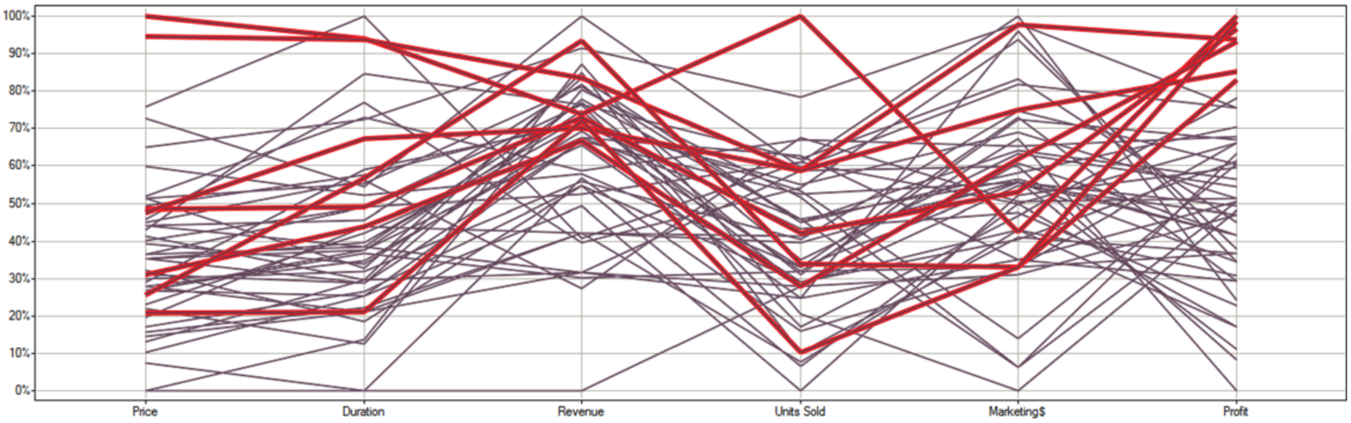


Figure 12.9. Created using Spotfire

It's now a bit easier to trace the paths of these seven lines without losing sight of the other data. In this case, the lines that aren't selected are still a little too distracting. If the unselected lines were lighter, this would probably work just fine. Or we can rely on the other method for focusing on the seven high-profit products only: filtering out all products but those with high profits, as I've done in the next graph below. Now that only seven lines remain, it's easy to check for a predominant multivariate pattern associated with highly profitable products.

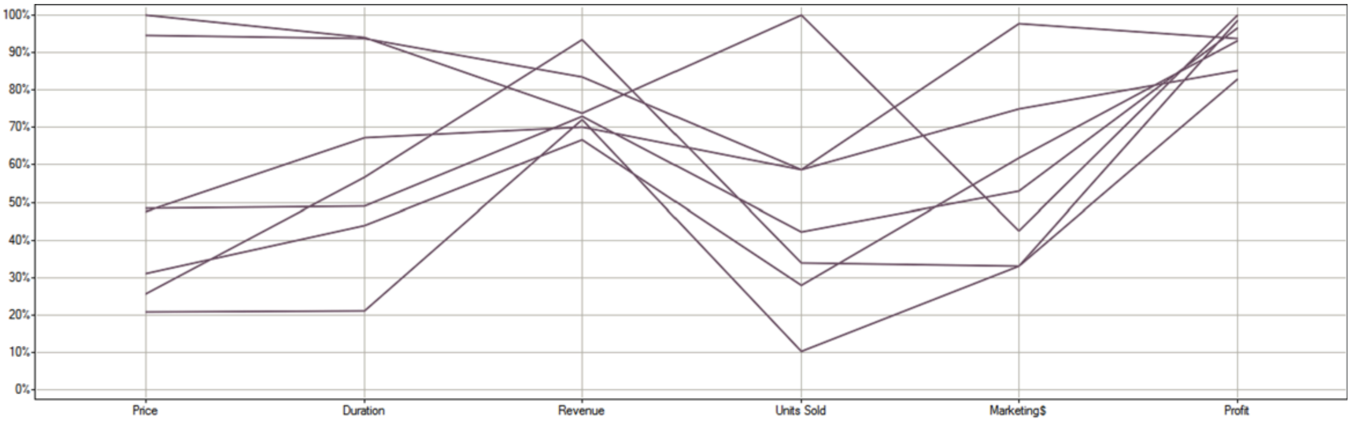


Figure 12.10. Created using Spotfire

We can now see that products with high profits exhibit a great deal of diversity in their overall multivariable profiles. A few significant patterns can be discerned, however. For instance, all products with high profits also have high revenues, which is no surprise. Also, in no case do marketing expenses fall below 30%. If we're hoping to produce high profits, the most we can say based on these particular variables is that we should try to generate high revenues and always invest more than a little in marketing.

To use parallel coordinates plots to best advantage, we need software that offers this kind of display along with good filtering and brushing functionality, but we can so something similar—identifying exceptions and predominant patterns in a set of multivariate data—using a tool as simple as Excel. By using a line graph with a line for each item that connects its value for each variable, you can create a display that looks like a parallel coordinates plot. Because a normal line graph uses the same quantitative scale for all values, however, you must normalize the quantitative scale associated with each variable so they are all the same. This can be done by converting each value of a particular variable to a percentage ranging from 0 to 100% so that all the variables share a common scale.

Multivariate Analysis Techniques and Best Practices

Visual data analysis works best in some cases when assisted by behind-the-scenes data crunching, rather than relying on our eyes alone. When we want to analyze multivariate data, computers can assist by performing the following two tasks:

- Ranking items by similarity
- Clustering items by similarity

Ranking Items by Similarity

Sometimes we examine multivariate data to find items that come close to matching a particular multivariate profile. For example, we might be interested in finding, among a pool of many thousand products, those with relatively low prices that have been on the market for an average amount of time, generate high revenues but have sold a relatively small number of units, and have lower-than-average marketing expenses yet high profits. It would be useful if we could describe this pattern and ask the software to find all products that come close to it. To illustrate how some software products do this, I'll use the same data set as above. The product that I'm using to illustrate this process, *Spotfire DecisionSite*, can search for patterns based on a multivariate profile that actually exists in the data set (including one that you enter specifically for this purpose), so I've selected one that comes close to the pattern described above, which I've highlighted in the example below.

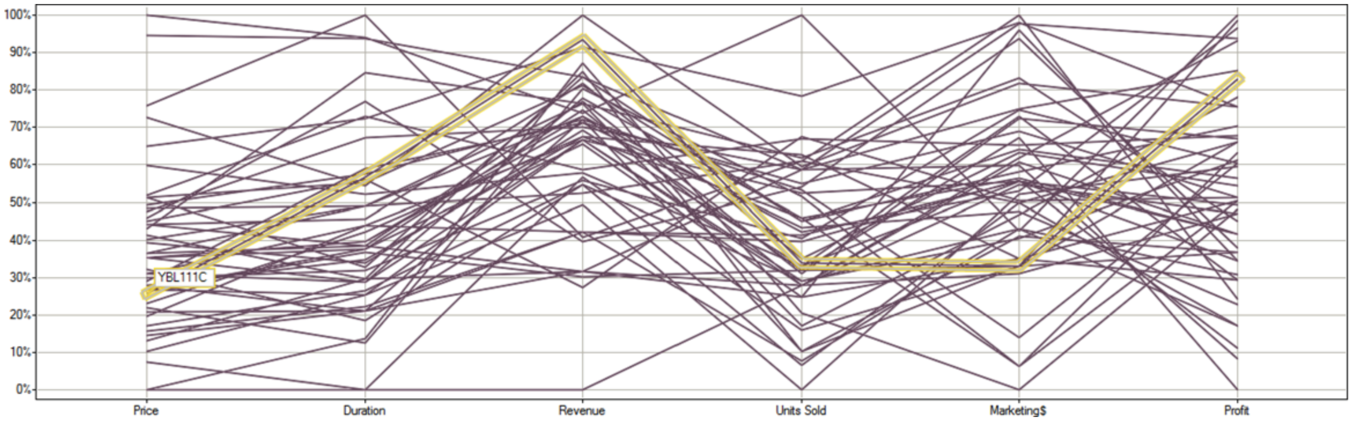
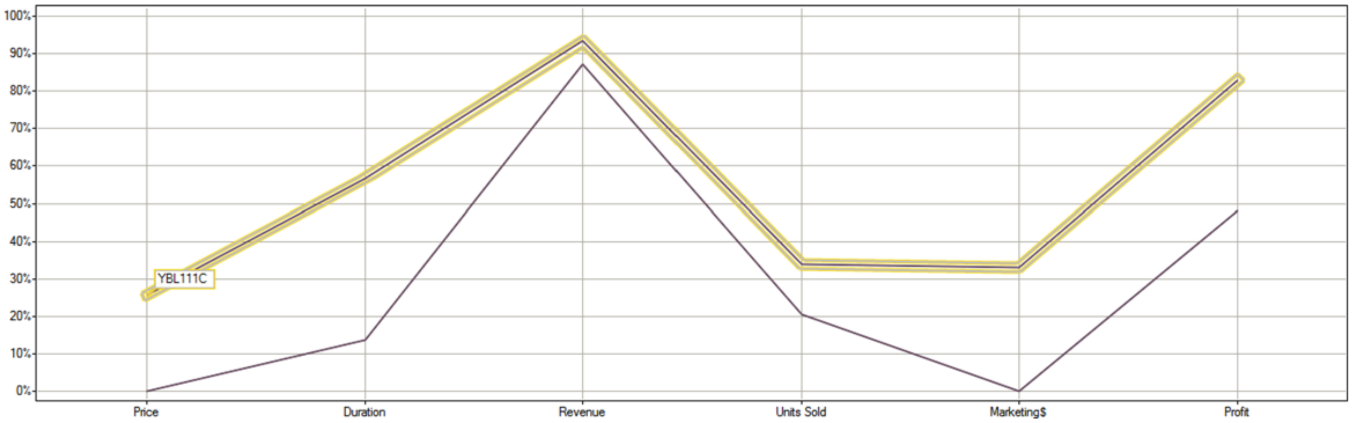


Figure 12.11. Created using Spotfire

After selecting this particular profile, I took advantage of Spotfire's pattern-detection functionality and asked it to search for similar profiles. In response, it automatically ranked every one of the 49 products from lowest to highest, based on how closely each corresponds to the selected profile. It also assigned a correlation coefficient value ranging from -1 (perfect negative correlation) to +1 (perfect positive correlation) to each. In the example below, I filtered out all but the selected profile (the highlighted line) and the one that is most similar to it.

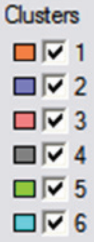


12.12. Created using Spotfire

The overall correlation coefficient of these two profiles is 0.90, which indicates a high positive correlation. When we're dealing with hundreds or perhaps thousands of items, our eyes aren't the best tools for spotting items that fit a particular multivariate profile. On such occasions, it makes sense to rely on computers to do what they do well—rapid and accurate calculations—and then let our eyes take over to examine the results.

Clustering Items by Similarity

Clustering is the process of segmenting data into groups whose items share a similar feature or features. In the example that we've been examining, similarity is determined by how much products are alike based on the total set of variables that we're examining (price, duration, and so on). Once again, computers are well equipped to do this work using a statistical clustering algorithm, which would be difficult and time consuming to do using our eyes. I asked the software to cluster the products for me, and it produced the six groups that appear below, each with its own color:



Clusters

🟧 ☑ 1
🟦 ☑ 2
🟥 ☑ 3
⬛ ☑ 4
🟩 ☑ 5
🟦 ☑ 6

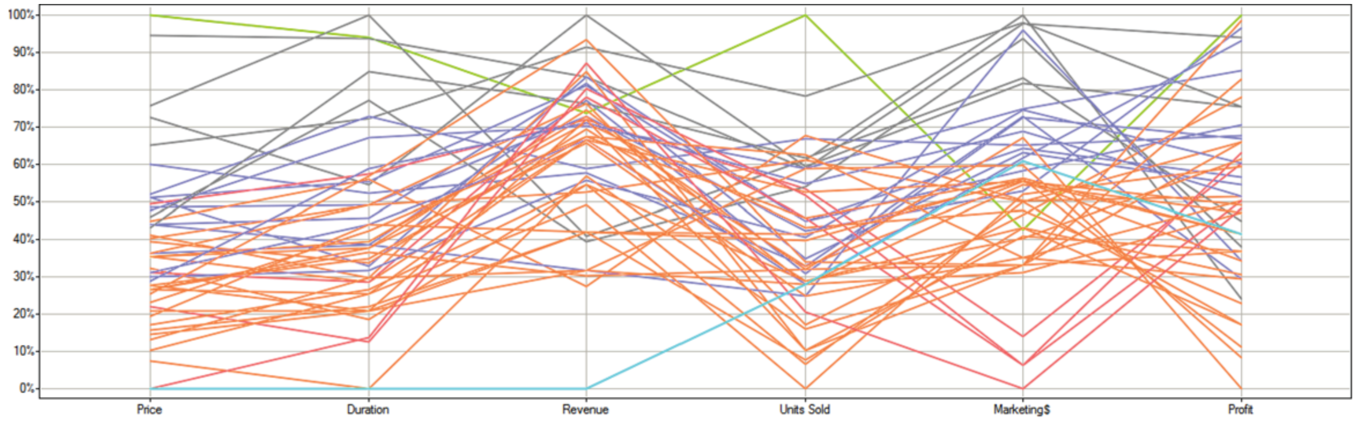


Figure 12.13. Created using Spotfire

Even though the groups are distinguished by color, it's hard to see their similarities in the midst of this visual clutter. To make it easier, I asked the software to separate the groups using a trellis display, with one group one per graph.

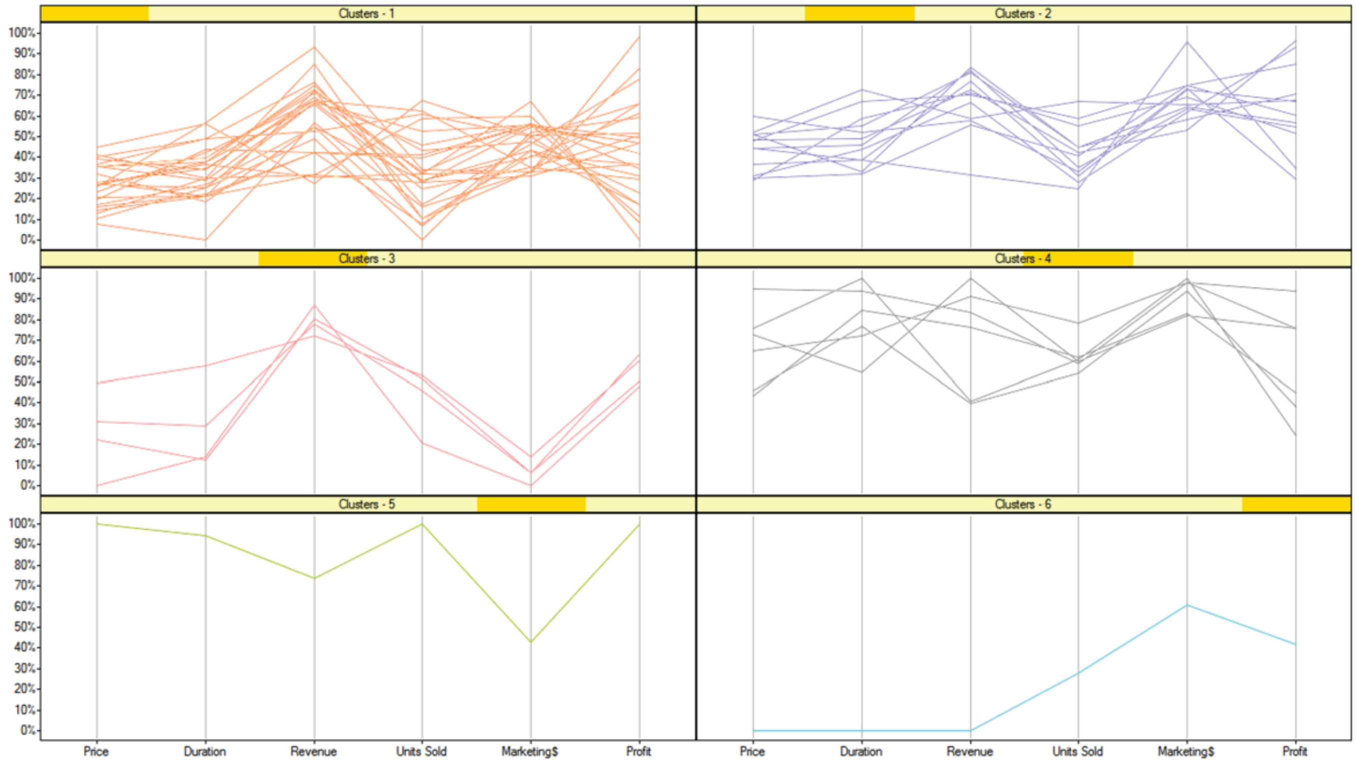


Figure 12.14. Created using Spotfire

Two of the products have such unique multivariate profiles (clusters 5 and 6), they've been placed in groups of their own. The product in cluster 5 has the highest profits of all. A total of 25 products were placed in cluster 1. The fact that this many products were this much alike wasn't obvious when viewing all the products together. Notice how similar to one another the products in the third group appear to be. The consistent midrange profits and low marketing expenses make these products worth further examination.

Statistical clustering algorithms are mathematically complicated. I won't pretend to be able to judge the relative merits of the many that are available. I trust that each is capable of doing the job and recommend that you use whatever method your software supports. If several similarity-matching methods are available, stick with the default method unless you've taken the time to learn their differences and the circumstances that make one better than the others.

This introduction to multivariate analysis and to parallel coordinates plots in particular is brief and designed to whet your appetite just enough to interest you in learning more. I hope you've recognized the rich potential of multivariate analysis.