perceptual edge

Multivariate Analysis Using Heatmaps

Stephen Few
October 10, 2006

This is the third article in a series that began in July with the article entitled, "An Introduction to Visual Multivariate Analysis." Prior articles in this series have examined how *table lens* and *parallel coordinates* displays can be used to explore and analyze multivariate information. In this article, I describe the use of *multivariate heatmap matrices*.

In general, the term *heatmap* refers to any display that uses color to represent quantitative data. We are all familiar with heatmaps in the form of weather maps, which use color to encode values such as temperature or rainfall. Heatmaps also come in forms other than geographical maps. When heatmaps are used to encode multivariate data—several variables that measure different aspects of some set of entities (for example, customers, countries, or products)—they are usually structured as a matrix of columns and rows. Figure 1 is a multivariate heatmap matrix, which displays a separate employee per row (the entities) and a separate measure per column (the variables). In this case, the heatmap's purpose is to help us determine what factors most influence employee job satisfaction, which appears in the leftmost column labeled *Working Conditions*. Employees were asked to rate their working conditions as *Very Poor* (the lightest color), *Poor*, *Acceptable*, *Good*, or *Very Good* (the darkest color). Each of the other variables (*Salary*, etc.) has been encoded as a continuous range of grayscale colors, ranging from the lightest for the lowest value through the darkest for the highest value. By examining a single row, you can see a particular employee's complete multivariate profile. By scanning a column, you can see the complete set of values for a particular variable across all employees, such as the average number of hours they work per week (the third column).

Figure 1: A typical multivariate heatmap matrix.

Take some time before reading the next paragraph to examine this heatmap on your own. See if you can determine which of the five variables (salary, average hours per week, etc.) seem to correlate significantly to these employees' perception of their working conditions.

You may have noticed that three variables exhibit what appear to be significant correlations to these employees' perception of their working conditions: average hours per week, average percent travel, and self rating. Employees who work long hours tend to rate working conditions as poor and vice versa. Those who travel a great deal for the job also rate working conditions as poor. Those who rate their own performance as poor tend to rate their working conditions similarly. Some degree of correlation might also exist between employees' perception of working conditions and their salaries and their supervisor's ratings, but these correlations are not as clear-cut.

I created this heatmap using Spotfire DecisionSite, based on data in an Excel spreadsheet. Visual analysis software that supports the use of heatmaps typically provides several ways to interact with the data, such as sorting and filtering, which helps to bring meaningful patterns to light. Figure 2 shows more of the screen, revealing filter controls to the right of the heatmap in the form of sliders for quantitative variables and a series of checkboxes for working conditions.

Figure 2: Spotfire DecisionSite displaying a heatmap.

To confirm our observations about the correlations of these variables to employees' perception of working conditions, I've isolated those parts of the heatmap that display employees who rated their working conditions as either *Very Good* or *Very Poor* (see Figure 3).

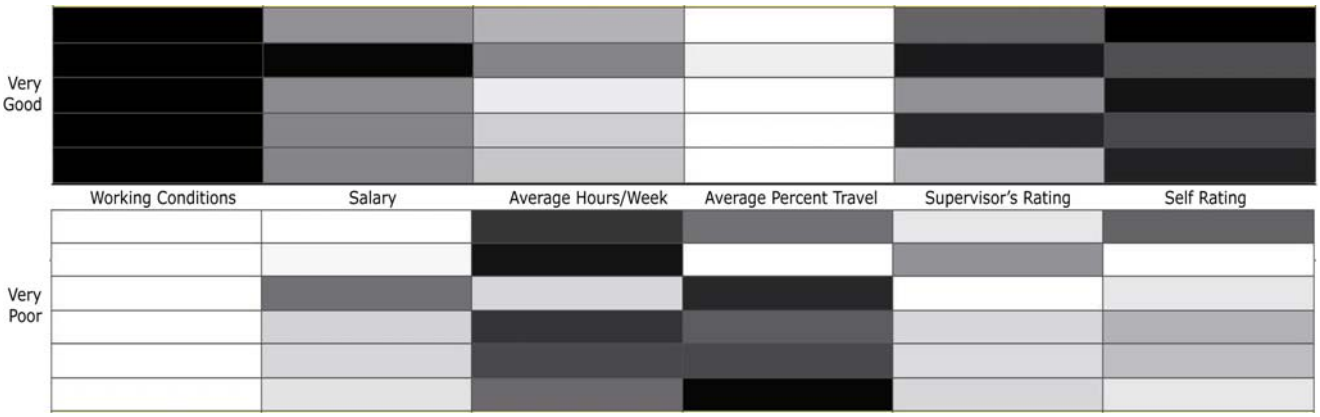

Figure 3: Heatmap display filtered down to only those employees who rated their working conditions as either *Very Good* or *Very Poor*.

This view allows us to observe, with less distraction, the fact that those who rate their working conditions as *Very Good* tend to work fewer hours per week, travel little, and rate themselves very highly, and that those who rate their working conditions as *Very Poor* tend to exhibit the opposite conditions. What also becomes clearer in this display is the fact that those at the extremes (*Very Good* and *Very Poor*) also exhibit significant correlations to salary and their supervisors' ratings as well, which is not as evident in the middle range of *Good*, *Acceptable*, and *Poor* working conditions.

Another way to test our observations involves sorting and filtering the data. In Figure 4, I've removed from view all employees who work less than 60 hours per week and sorted those that remain by the number of hours worked in ascending order, which you can see as a continuous gradient of lighter to darker gray in the third column from the left. Notice that for this set of employees, ratings of working conditions tend to be low, exhibited by mostly light shades of gray.

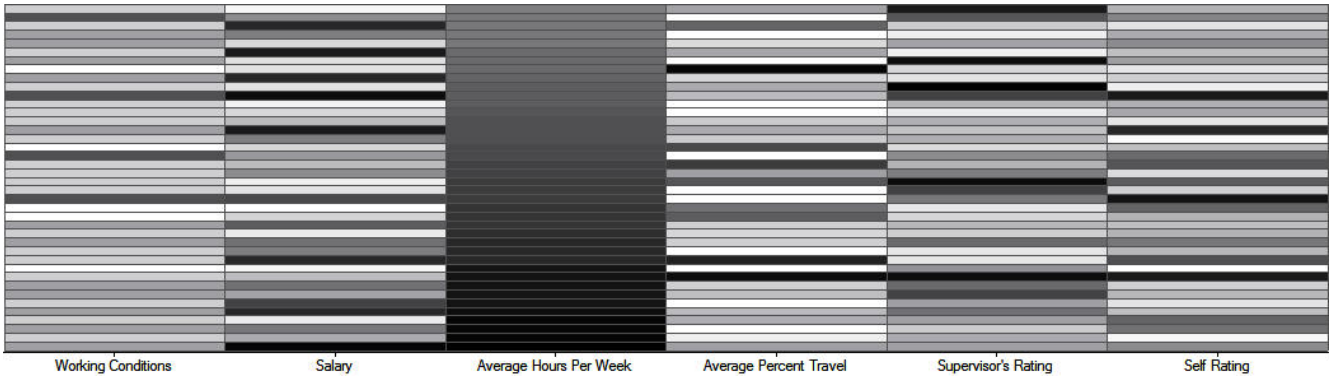

Figure 4: This heatmap only displays employees who work long hours.

In Figure 5, you see the results of a similar filtering and sorting operation, this time on the *Average Percent Travel* column, featuring only those who spend 50% or more of their time on the road. Once again, the ratings of working conditions tends to be on the low side, although not quite as strongly as what we saw related to long work hours.

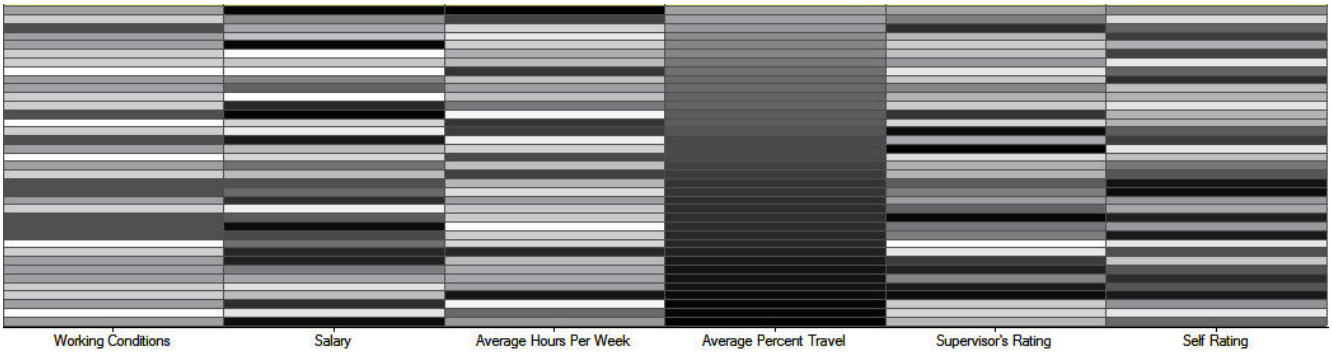

Figure 5: This heatmap only displays employees who spend 50% or more of their time traveling.

Multivariate heatmap matrices are often used by scientists to analyze data. Rather than using them by themselves, similar to the examples that I've shown, however, they tend to use heatmaps in combination with a complementary display called a *dendogram*. Despite the fancy name, a dendogram simply organizes the entities hierarchically, based on similar multivariate profiles, displayed as a tree structure. Figure 6 shows an example of the employee satisfaction data that we've been reviewing, this time as a combined dendograms and heatmap.

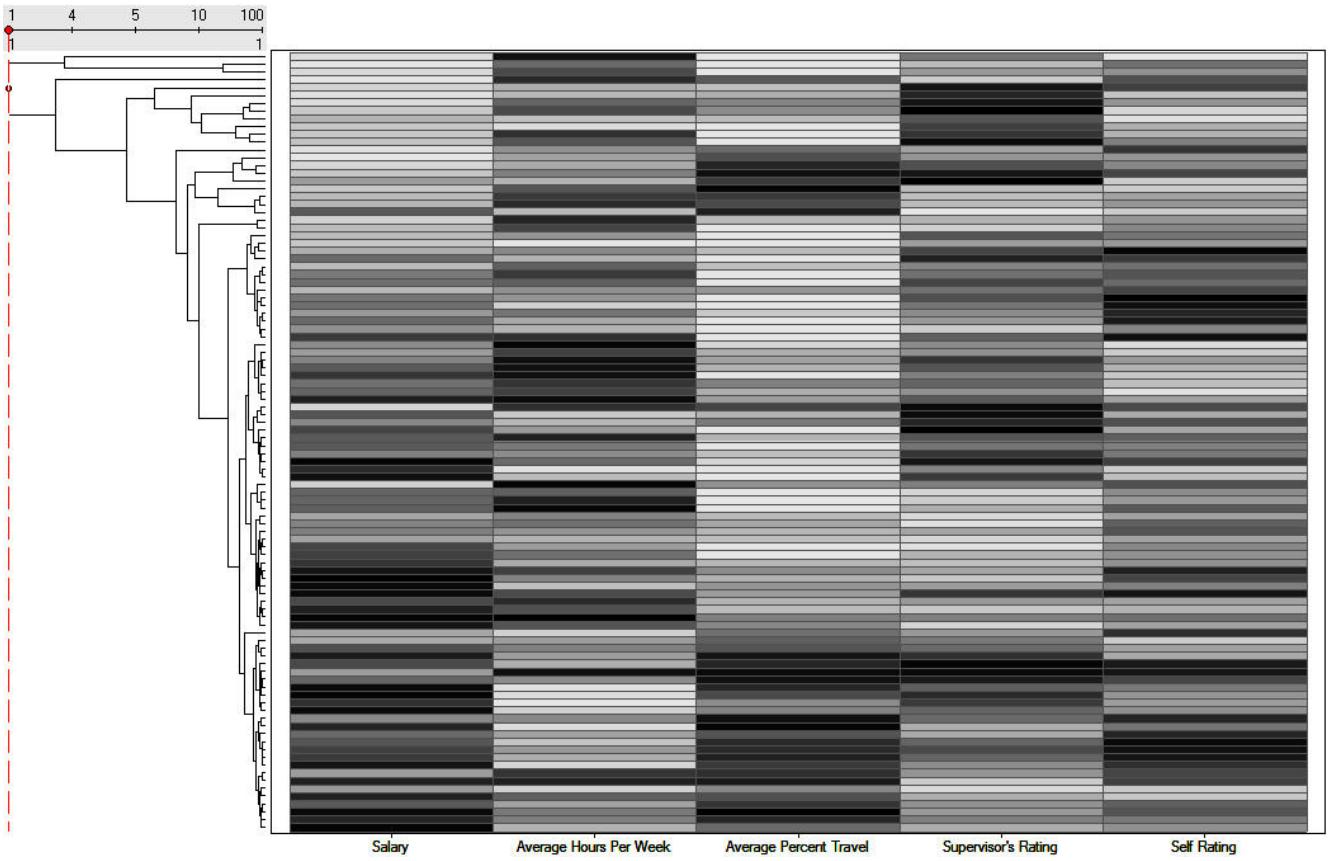

Figure 6: A heatmap that has been hierarchically organized by a dendogram on the left.

Employees who are linked together at the lower levels of the hierarchical tree (the smaller clusters on the right) are very similar in their multivariate profiles, and their similarities become more generalized and as you proceed up the hierarchy (the larger clusters on the left). Scientists who use these displays are statistically sophisticated and have spent a great deal of time learning to read them. There are many statistical means to choose from when determining how items should be clustered into similar profiles, which must be understood to use this kind of display effectively. This type of analysis is usually reserved for the statistical elite, not most of us who are simply trying to make sense of business information.

Whenever color is used to encode quantitative values, a few simple rules should be followed to make the display understandable. Primarily, you must understand that we don't perceive a rainbow of hues quantitatively. Which is greater: orange, blue, green, red, black, or purple? If you've memorized the color spectrum arranged in order of wavelength, with much practice you could use that knowledge to interpret differences in quantitative values that have been encoded across a spectrum of hues, but this isn't easy and certainly isn't intuitive. Two basic methods of encoding continuous quantitative values work well for heatmaps. The first, sometimes called a *sequential scale*, consists of a single hue expressed as a gradient that varies from light to dark or pale to fully saturated, which is what I've been using in the examples so far. We can easily and intuitively perceive a gradient of color intensity as a range of continuous quantitative values.

The other method, sometimes called a *diverging scale*, consists of two hues, with a neutral color such as gray in between, with each hue expressed as a range of intensity from light to dark or pale to fully saturated. This method is useful when a range of quantitative values is

logically divided into two groups, such as negative and positive values with zero in the middle (for example, positive and negative profits). Figure 7 illustrates a diverging scale that could be used on a heatmap.

Figure 7: This diverging color scale could be used to encode negative values in red and positive values in blue.

Logical breakpoints in a range of quantitative values other than negative and positive numbers can be encoded using a diverging scale, such as the number of manufacturing defects that range above or below average or the performance of salespeople that exceeds or falls below quota.

How well does a heatmap work for multivariate analysis compared to parallel coordinates, which we examined last month? I haven't tried to compare the effectiveness of these two approaches in an empirical manner, so I cannot state a firm opinion, but I can make a few tentative observations. Figure 8 shows the same set of data that we've been examining, now displayed as parallel coordinates.

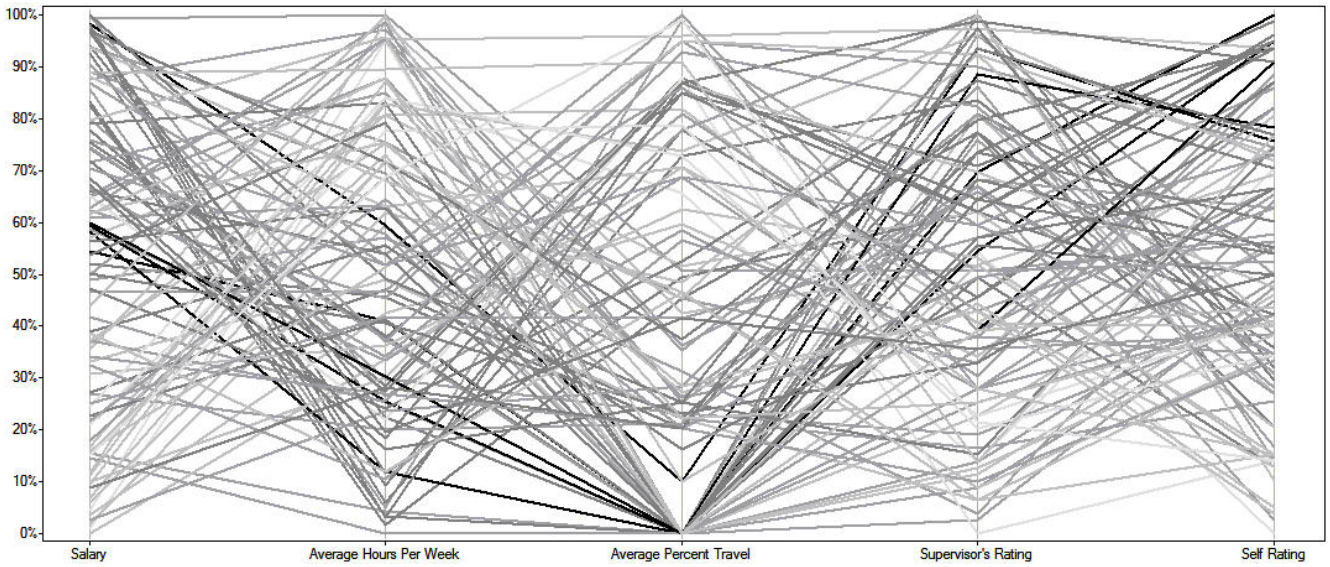

Figure 8: A parallel coordinates display of the employee satisfaction data.

In this display, each employee's rating of working conditions has been encoded as the color intensity of the line, ranging from light for the lowest rating through dark for the highest rating. For each of the quantitative variables (the five vertical axes), the position where the line intersects the axis corresponds to the value, with the bottom-most point representing the lowest value and the top-most point representing the highest. How do you think the parallel coordinates display compares to the heatmap?

Here are two observations regarding this comparison:

- Heatmaps don't suffer from the problem of occlusion (objects hiding behind and being obscured by other objects), which can make parallel coordinates look cluttered where many lines appear in the same vicinity. Each measure in a heatmap, however, resides in its own cell in the matrix, so there is never any occlusion.

- Multivariate profiles exhibited as a series of colors (such as in a single row in our sample heatmap) are not as easy to perceive and remember as the single pattern of ups and downs formed by lines in parallel coordinates display.

The problem of occlusion in a parallel coordinates display can be reduced by dividing the data into a series of separate parallel coordinates graphs, such as shown in Figure 9, which separates each rating of working conditions into its own graph.

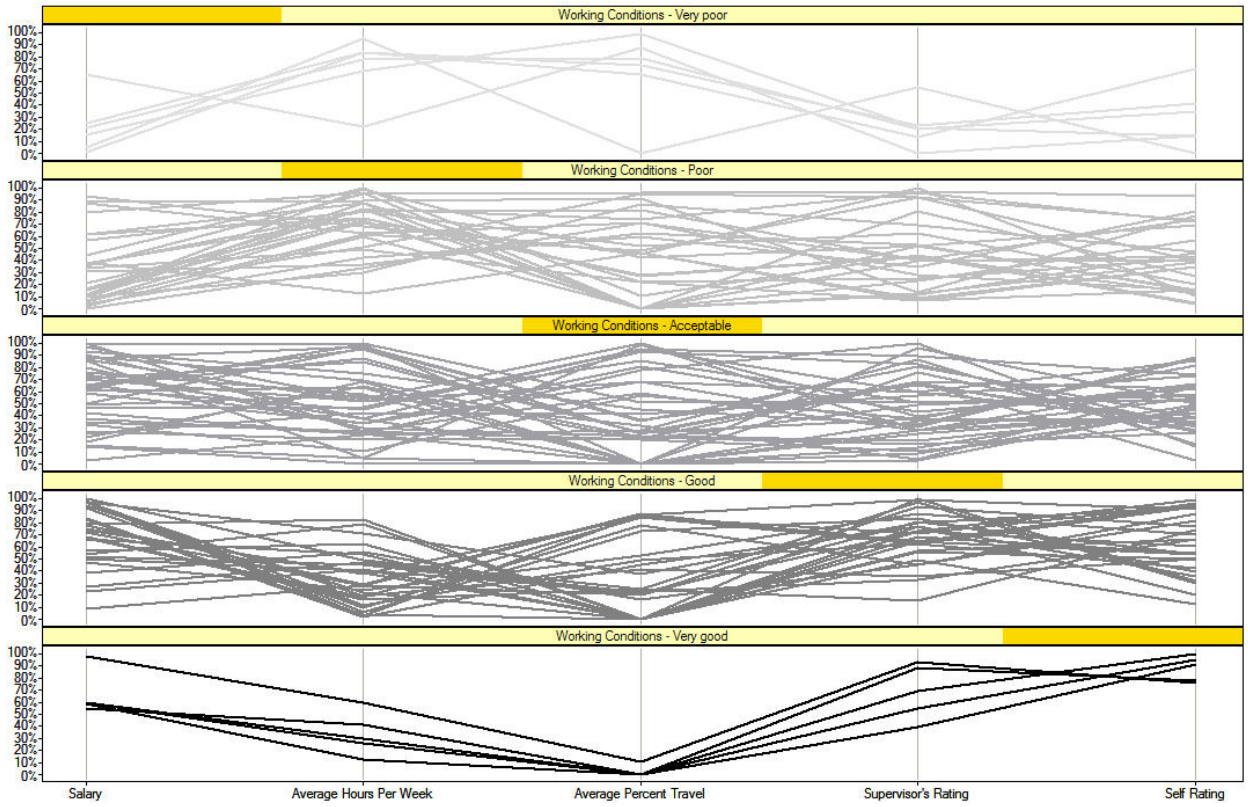

Figure 9: A series of parallel coordinates graphs—one for each rating of working conditions.

Displayed in this manner, the one advantage that a heatmap seems to offer when compared to parallel coordinates is considerably diminished, resulting in a display that exhibits clearer multivariate patterns that can be more easily compared. Given empirical evidence to the contrary or examples of heatmap displays that offer advantages unfamiliar to me, I could certainly be persuaded to change my opinion, but for now I see an advantage in using parallel coordinates for the analysis of multivariate business data by people who are not highly trained statisticians.

In next month's article in this series, we will examine an approach to multivariate analysis that encodes a set of variables in the form of a *glyph*—an object that consists of multiple parts,

each of which represents a separate variable combined to display an entire multivariate profile.

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.