

6 ANALYTICAL PATTERNS

Quantitative analysis involves examining relationships among values. In this book, I've categorized data analysis based on the nature of the relationship that's being examined. Analyzing these relationships requires us to search for particular patterns in data and to use particular analytical techniques. Part II, which begins with the next chapter, explains how to analyze each of these types of relationships, one per chapter:

- Time-series
- Ranking and Part-to-Whole
- Deviation
- Distribution
- Correlation
- Multivariate

This chapter gives an overview of different ways that we can represent data visually before we begin to examine the relationships above in depth in the next few chapters.

When any of the relationships above are represented properly in visual form, we can see particular patterns and analyze them to make sense of the data. To prime our eyes for pattern perception before diving into specific types of analysis, we'll take some time now to think about patterns that are meaningful in several types of analysis.

Remember that in *Chapter 3: Visual Perception and Information Visualization*, I explained that our visual sense receptors are highly tuned to respond to particular low-level characteristics of objects called pre-attentive attributes. These basic attributes of form, color, position, and motion can be used to display abstract data in ways that are rapidly perceptible and easily graspable. When we look at a properly designed graph, we can spot patterns that reveal what the information means. Although graphs inform us differently than spoken or written words, both involve language: one is visual and the other verbal. Similar to verbal language, visual displays involve semantics (meanings) and syntax (rules of structure). Letters of the alphabet are the basic units of verbal language, which we combine to form words and sentences according to rules of syntax that enable us to effectively communicate the meanings we intend. In the same way, simple objects such as points, lines, bars, and boxes are basic units of visual language, which are combined in particular ways according to rules of perception to reveal quantitative meaning.

Guidelines for Representing Quantitative Data

Particular visual objects are best suited to communicating particular quantitative relationships for particular purposes. It helps to know the strengths and weaknesses of each type of visual object so that we can choose the type of graph that will best help us find what we're looking for, examine it in the way will most likely to lead to understanding, and, when necessary, communicate the information to others.

Bars

When you look at two objects like the two dark rectangles below, what do you notice and what meanings come to mind?



Figure 6.1

What likely stands out most prominently to most of us is the difference in their heights. This difference invites us to notice that one bar is taller than the other. This is what bars are especially good at: displaying differences in magnitude and making it easy for us to compare those differences. Also, because bars have such great visual weight and independence from one another, like great columns rising into the sky, they emphasize the individuality of each value.

The following graph makes it quite easy to see planned versus actual sales in each region as distinct and to compare magnitudes with accuracy and little effort. It is especially easy to compare planned versus actual sales because they are next to one another. In other words, this graph, by the way it has been designed, guides us to make that particular comparison, just as the choice and arrangement of words in a spoken or written sentence points us toward certain meanings and interpretations and away from others. This is what I meant previously when I said that we must understand and honor a visual equivalent of vocabulary and syntax—the rules of perception—when we use graphs.

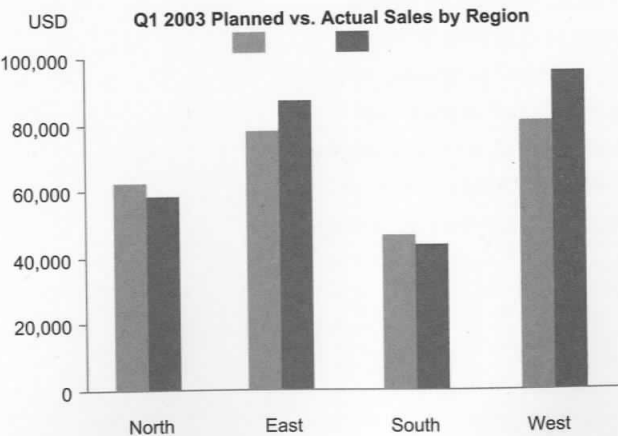


Figure 6.2

We'll usually choose bar graphs when we want to emphasize the individuality of values and compare their magnitudes.

Boxes

When you look at objects like the two subdivided rectangles below, what are you inclined to notice and what meanings come to mind?

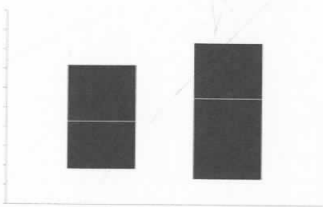


Figure 6.3

Although these rectangles are similar to bars, they don't share a common baseline, so we tend to notice the differences between the positions of their tops and their bottoms, the difference between the horizontal lines that divide them, and the difference between their total lengths. This is precisely what these rectangles are designed to help us do. They are called boxes, and the graphs in which they are used are called box plots. Each box represents the distribution of an entire set of values: the bottom represents the lowest value, the top represents the highest value, and the length represents the full spread of the values from lowest to highest. The mark that divides the box into two sections, in this case a light line, indicates the center of the distribution, usually the median or mean value. A central measure (also called an average) gives us a single number that we can use to summarize an entire set of values. Notice how your eyes are encouraged to observe and compare the different positions of the centers of these boxes, and how the difference in the position of the center lines conveys that, on average, the values represented by the box on the right are higher than those on the left. The graph below illustrates the usefulness of box plots. In this case, the graph can be used to compare the distributions of salaries for five years and to see how they changed from year to year.

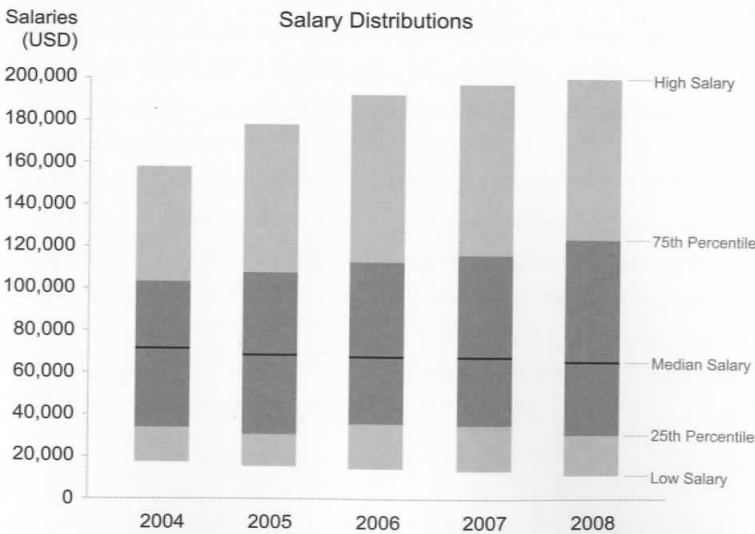


Figure 6.4

The box plot on the previous page tells the story of how employees are compensated in an organization, based on five values that summarize each year's distribution: the highest, lowest, and middle values as well as the point at and above which the top 25% of salaries fall (the 75th percentile), and the point at and below which the bottom 25% of salaries fall (the 25th percentile). The example below displays the same exact salary distributions in a way that is more typical of how box plots are usually drawn.

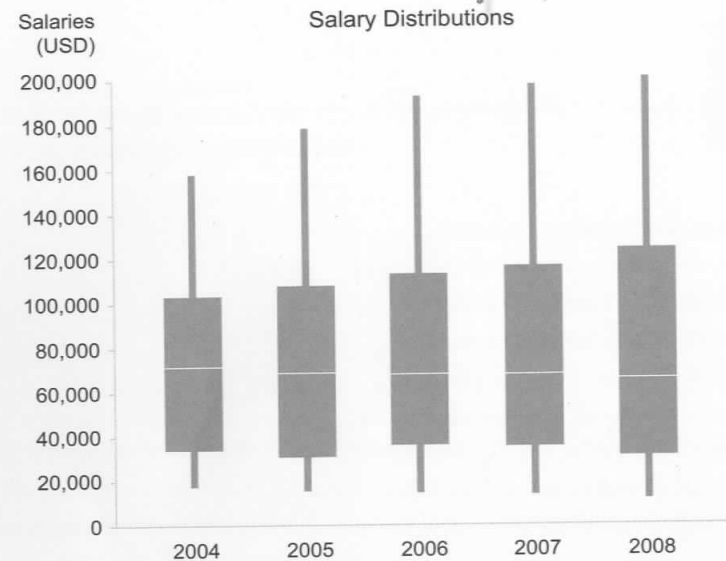


Figure 6.5

We'll take the time to learn more about box plots in *Chapter 10: Distribution Analysis*. If you haven't used them before, you'll find that they'll become familiar in no time.

Lines

When you see an object like the line below, what does it suggest?



Figure 6.6

This particular line, which angles upwards from left to right, suggests an increase, something moving upward. Lines do a great job of showing the shape of change from one value to the next, especially change through time.

The strength of lines is their ability to emphasize the overall trend and specific patterns of change in a set of values. The following graph tells a vivid story of how sales changed throughout the year.

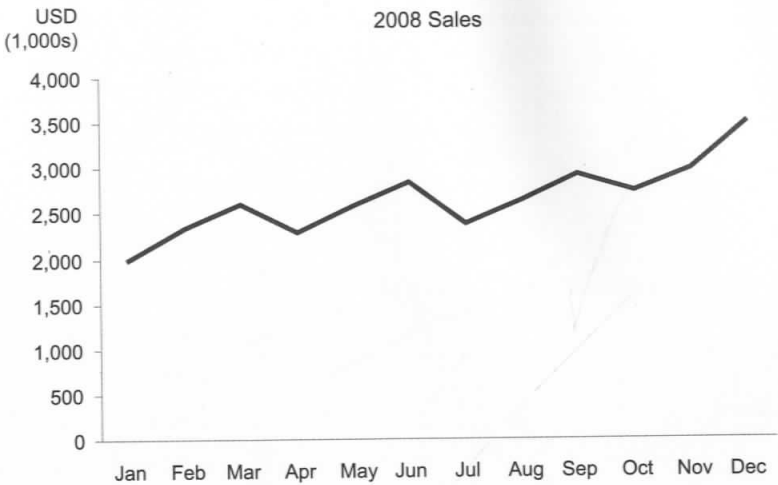


Figure 6.7

Points

When you see points scattered about, as illustrated below, what features attract your attention?

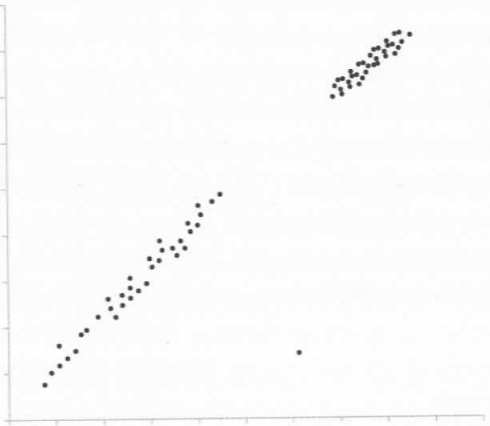


Figure 6.8

Points, such as those in a scatterplot, encourage us to notice patterns such as clusters, linear or curved arrangements, gaps, and points that appear isolated from the majority. These are precisely the patterns that are meaningful in correlation relationships, which is what scatterplots were designed to display.

Perhaps the greatest strength of points is the ability of each to encode two quantitative values: one based on its horizontal position and one based on its vertical position. Looking at the scatterplot on the following page, which displays a potential correlation between advertisements and resulting sales orders (two quantitative variables), it is hard to imagine any object other than a point that could do the job as well. Bars certainly wouldn't work when the X-axis and Y-axis both host quantitative scales, and if we connected the points with a line, the result would be a meaningless string of spaghetti.

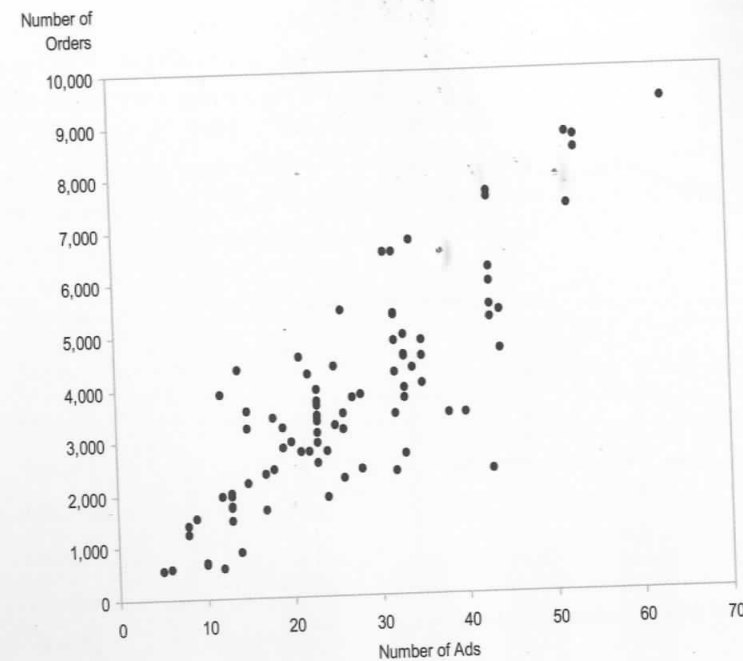


Figure 6.9

Points can also be used as a substitute for bars when there is an advantage to narrowing the quantitative scale so that zero is no longer included. Remember that, when bars are used, the quantitative scale must include zero as the baseline for the bars because otherwise the lengths of the bars will not accurately encode their values. (See the section on comparing and contrasting in *Chapter 4: Analytical Interaction and Navigation* for a more detailed explanation of this issue). In the bar graph below, all of the values fall between 62% and 83%. Most of each bar's length doesn't tell us much because we are mostly concerned with the differences between the values, and they all fall within a relatively narrow range at the right.

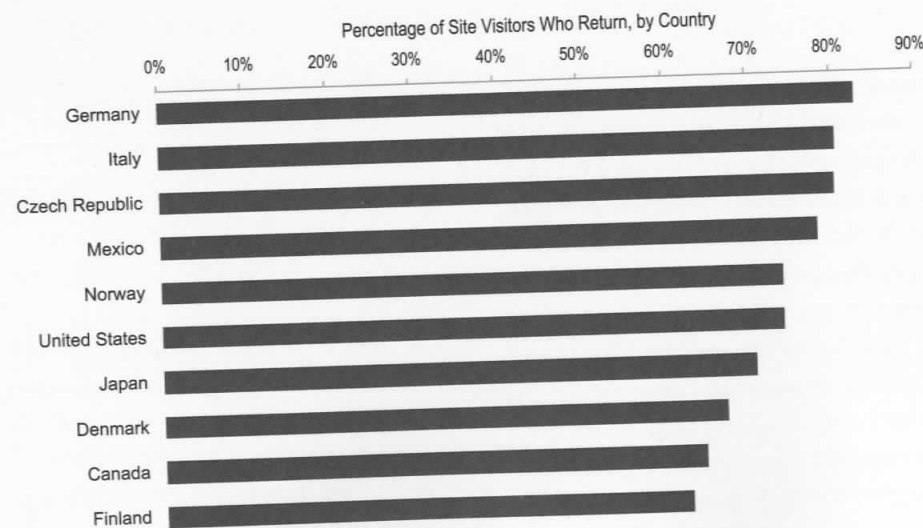


Figure 6.10

If we want to examine these differences more clearly, we can't just narrow the scale to begin around 60% because then the bars' lengths would no longer accurately encode the values. We could narrow the scale, however, if we replace the bars with points to create a dot plot, as shown below.

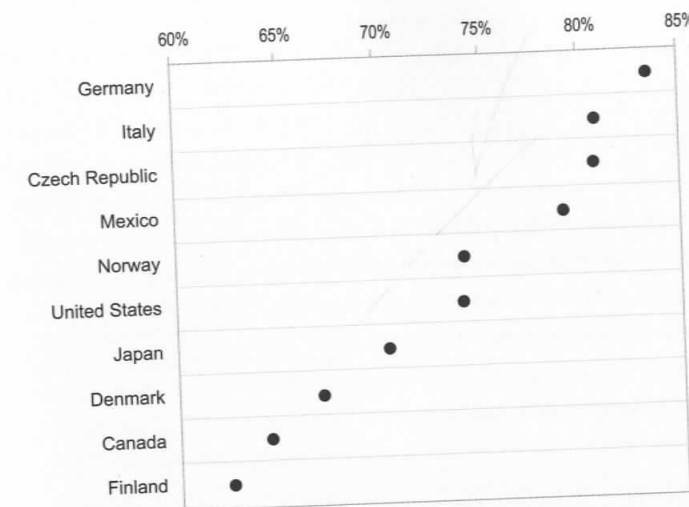


Figure 6.11

These points encode values based on their horizontal position in relation to the quantitative scale along the top. We are no longer comparing the lengths of objects, so the elimination of zero from the scale does not create the same perceptual problem that would have been created with bars.

Points and lines can be used together in a line graph to clearly mark the actual positions of values along the line. This is especially helpful when a graph displays more than one line, and we need to compare the magnitudes of values on different lines. For example, in the example below, if we want to compare domestic and international sales in the month of June, the points make it easier for our eyes to focus on the exact position on the line where the comparison should be made. When we primarily want to see the shape of change through time but secondarily also want to make magnitude comparisons between the lines, a line graph with data points works well.

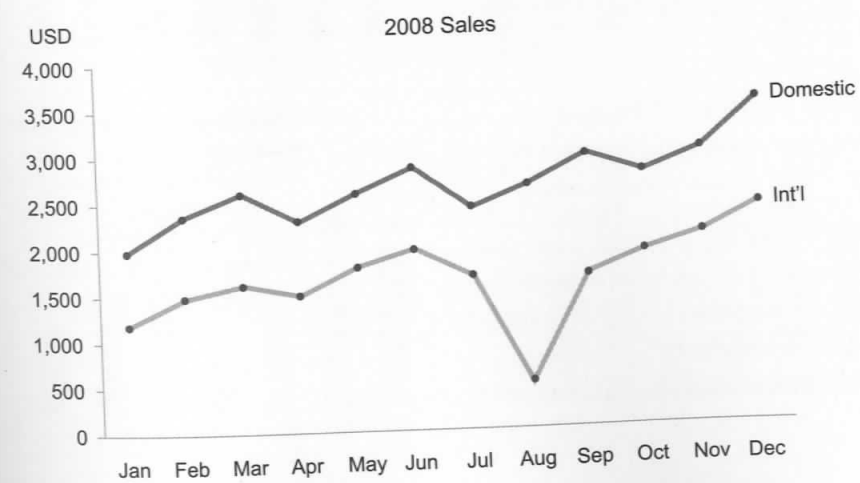


Figure 6.12

In the Smooth and in the Rough

Borrowing terms from statisticians, we can speak of data as falling into two categories: *in the smooth* and *in the rough*. Values that are relatively typical of the set as a whole reside in the smooth. Technically, *smooth* is just another term for “fit,” that is a line or curve that fits the shape of the data (also known as a trend line), which we use to describe the overall pattern in the data. Atypical values, those that fall outside the normal range and therefore cannot be described by the smooth, are said to reside in the rough. The total set of data therefore equals the smooth plus the rough.

The smooth is the underlying, simplified structure of a set of observations. It may be represented by a straight line describing the relationship between two variables or by a curve describing the distribution of a single variable, but in either case the smooth is an important feature of the data. It is the general shape of a distribution or the general shape of a relationship. It is the regularity or pattern in the data...What is left behind is the rough, the deviations from the smooth.¹

I’ll usually use the term *exceptions* to refer to abnormal values in a set of data. Exceptions are sometimes called *outliers*. Technically, the terms exception and outlier differ slightly in meaning. Both, however, are worth examining. A value is an exception whenever it falls outside defined standards, expectations, or any other definition of normal. Outlier, by contrast, is a statistical term that refers to values that fall outside the norm based on a statistical calculation, such as anything beyond three standard deviations from the mean.

An outlier is a value which lies outside the normal range of the data, i.e., lies well above or well below most, or even all, of the other values... It is difficult to say at just what point a value becomes an outlier since much depends upon its relationship to the rest of the data and the use for which the data is intended. One may want to identify and set aside outlying cases in order to concentrate on the bulk of the data, but, on the other hand, it may be the outliers themselves on which the analysis should be concentrated. For example, communities with abnormally low crime rates may be the most instructive ones.²

Outliers can...be described as data elements that deviate from other observations by so much that they arouse suspicion of being produced by a mechanism different than that which generated the other observations.³

Whether we use a strict statistical method to identify true outliers or some other approach to identify exceptions, we must first define what is normal in a way that excludes only those values that are extraordinary. Every abnormal value, whether an exception or an outlier, can and ought to be explained. Something has caused these unusual values. There are always reasons and it’s up to us to find them.

1. *Exploratory Data Analysis*, Frederick Hartwig with Brian E. Dearing, Sage Publications, Inc., Thousand Oaks CA, 1979, pp. 10 and 11.

2. *Ibid.*, pp. 27 and 28.

3. “Summarization Techniques for Visualization of Large Multidimensional Datasets,” Sarat M. Kocherlakota, Christopher G. Healey, Technical Report TR-2005-35, North Carolina State University, 2005, p. 4.

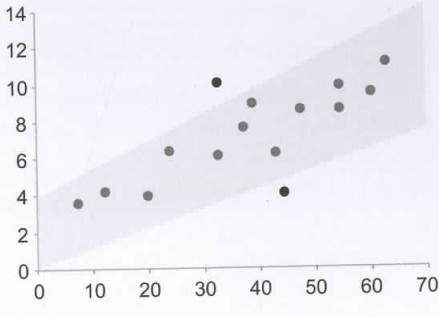
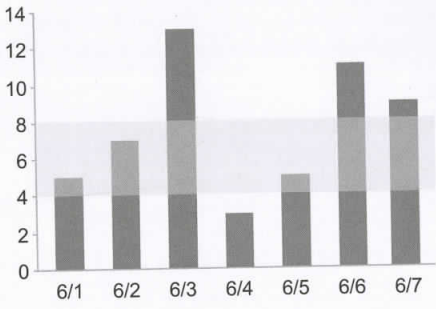
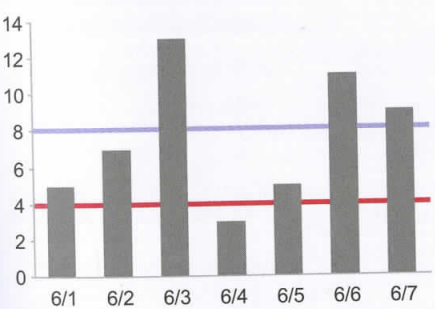
Values can fall outside the norm for three possible reasons:

- Errors
- Extraordinary events
- Extraordinary entities

Exceptions are often errors caused by inaccurate data entry, inaccurate measurements, or other mistakes. Some exceptions are not errors but are the result of extraordinary events: something happened, such as a storm, the loss of a key employee, or an exceptionally successful promotional campaign, that caused atypical behavior or results. And, finally, exceptions sometimes result from the behavior of a person, organization, or some other entity that itself falls outside of the norm, such as an order from a country that rarely makes purchases or a person with highly unusual tastes.

When examining information, we want spotting exceptions to be as easy as possible. Because memory is fallible, and working memory is in limited supply, we shouldn’t rely on memory to know the boundaries of normal. One of the great benefits of information visualization is the opportunity to offload cognitive workload to our eyes. Let’s not waste our brains on activities that our eyes can do faster and with much less effort. So in this example, it’s a good idea to display explicitly the boundaries that define the range of normal. The upper and lower boundaries of what we define as normal can be easily shown on a graph as reference lines or as reference areas of fill color that either define the range or define areas outside of the range. Standards that define what is acceptable can be displayed similarly.

Ranges of normal



Acceptable ranges (standards)

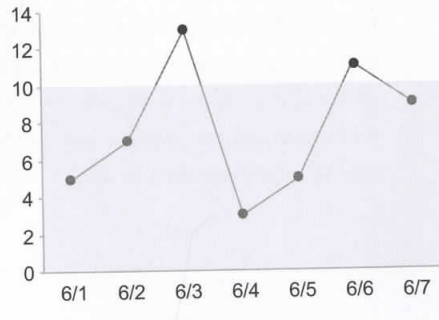
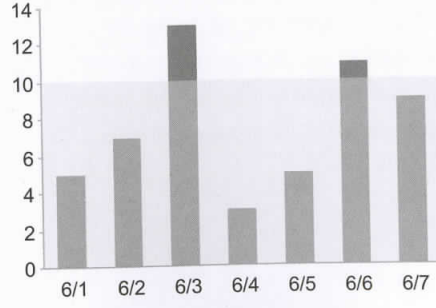
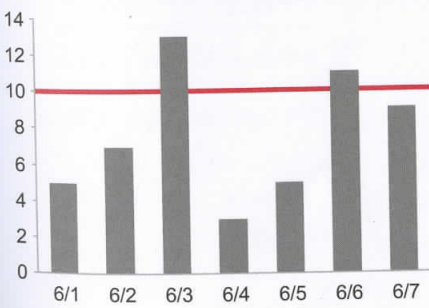


Figure 6.13

Pattern Examples

The number of unique visual patterns that exist in the world is virtually infinite. The number of patterns that represent meaningful quantitative information in 2-D graphs, however, is not. If we learn to recognize the patterns that are most meaningful in our data, we'll be able to spot them faster and more often, which will save us time and effort.

The example below no doubt appears overwhelmingly complex to most of us. But to someone who has been trained and developed expertise in reading this type of display, it isn't overwhelming at all.

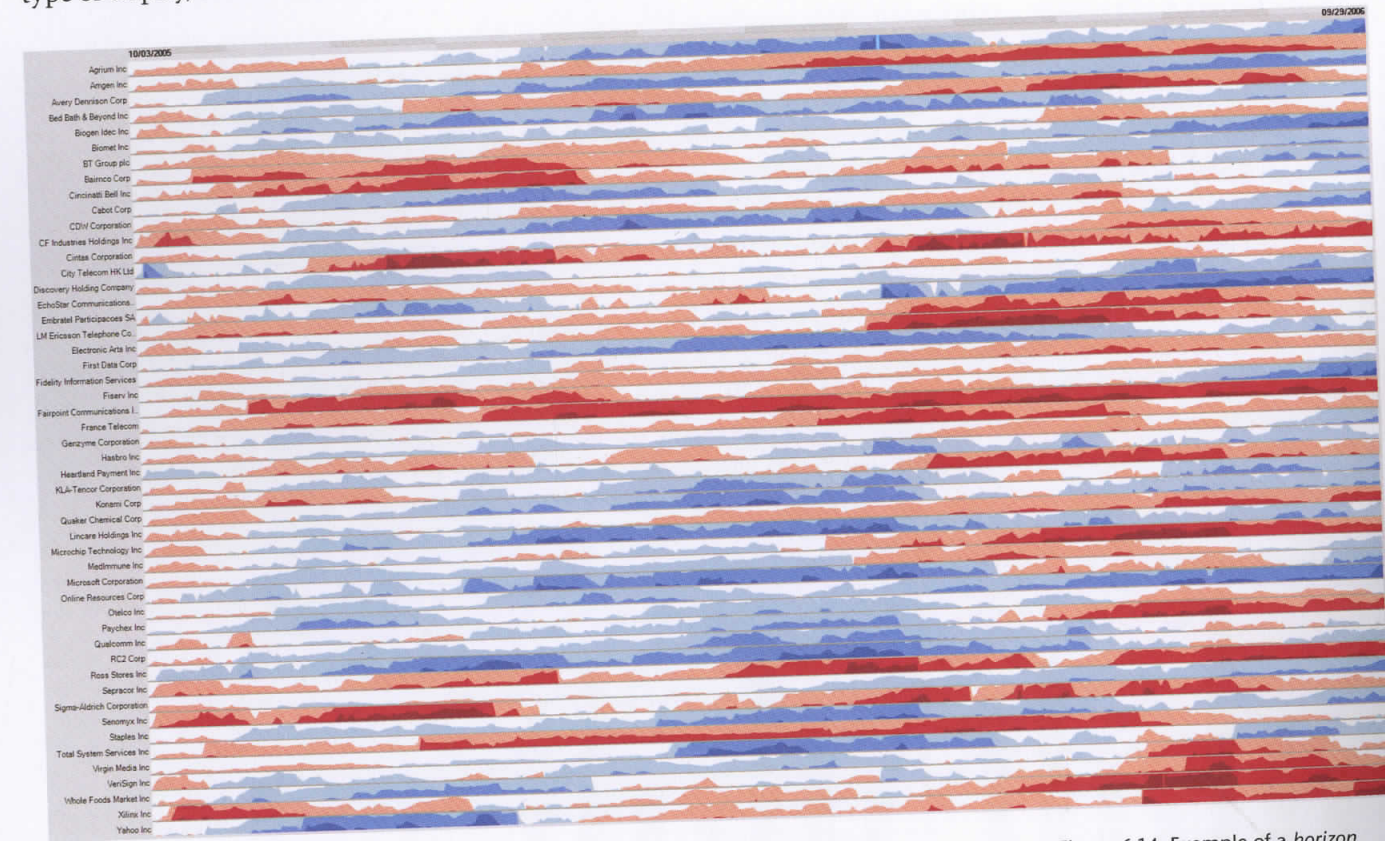


Figure 6.14. Example of a horizon graph, created using Panopticon Explorer

4. Information Visualization: Perception for Design, Second Edition, Colin Ware, Morgan Kaufmann Publishers, San Francisco CA, 2004, p. 209

Colin Ware states:

People can learn pattern-detection skills, although the ease of gaining these skills will depend on the specific nature of the patterns involved. Experts do indeed have special expertise. The radiologist interpreting an X-ray, the meteorologist interpreting radar, and the statistician interpreting a scatter plot will each bring a differently tuned visual system to bear on his or her particular problem. People who work with visualizations must learn the skill of seeing patterns in data.⁴

To an expert, much of what appears in the display isn't important; it's visual noise from which the meanings that matter can be easily and rapidly extracted. As visual data analysts, we must learn to separate the signal from the noise.

A number of basic patterns are almost always meaningful when they appear in graphs. They're not always relevant to the task at hand, so we won't always attend to them, but it's useful to hone our skills to easily spot them.

While looking at the blank graph below, try to imagine some of the meaningful patterns that might be formed in it by points, lines, bars, and boxes. Think about your own work, the data that you analyze, and call to mind patterns that catch your attention (or ought to) when present. Take a minute to list or draw examples of a few right now.

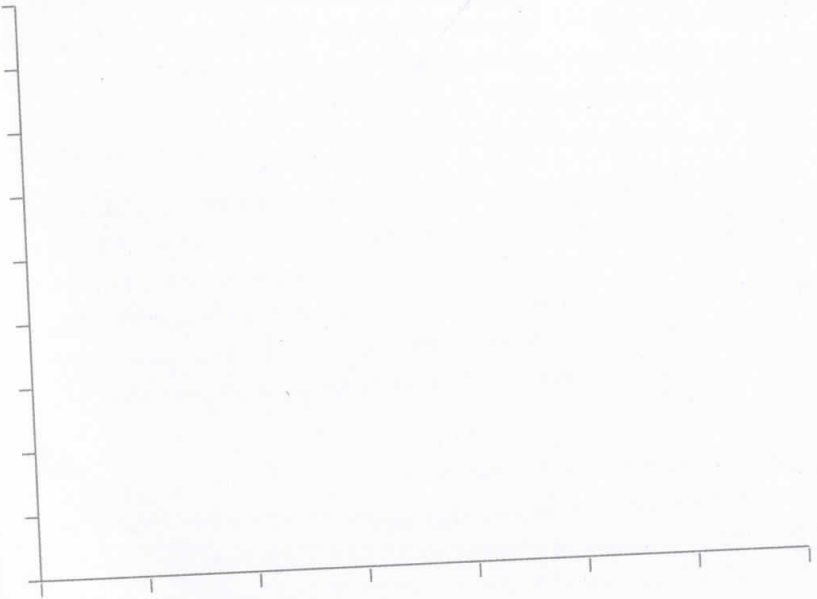
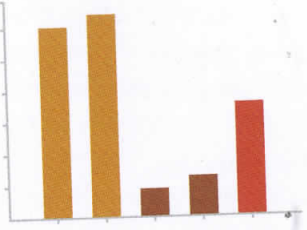
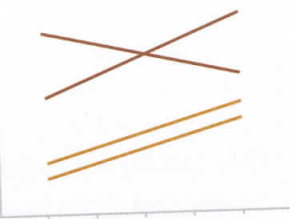
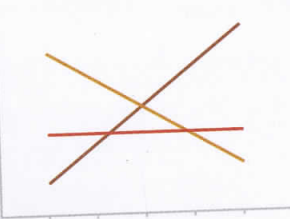
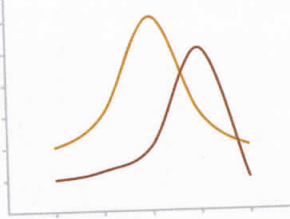

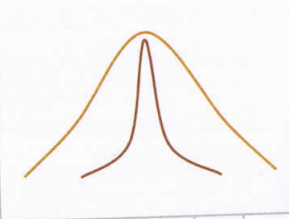

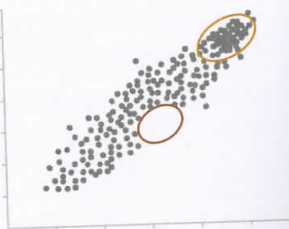
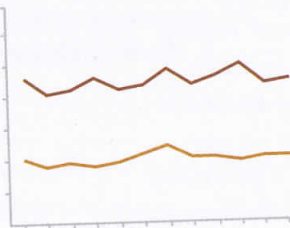
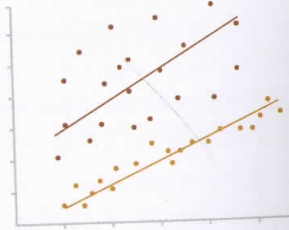

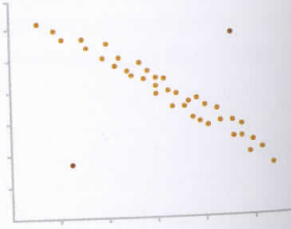


Figure 6.15

It helps to bring patterns to mind and understand what they mean when we spot them in various contexts. Doing this primes our perceptual faculties, sensitizing them to particular patterns, which helps us spot them more readily.

On the next page are examples of several patterns that are worth noticing when they show up in our data. Others might come to mind that are specific to your work and the kinds of data you encounter but I'm focusing here on patterns that are commonly found in data from lots of different types of businesses and other sources. This is by no means an exhaustive list, but we're likely to run across these patterns often. Part II presents more information on patterns as each chapter lists the specific patterns that apply to the type of analysis discussed in that chapter.

Pattern	Example	Pattern	Example
High, low, and in between		Non-intersecting and intersecting	
Going up, going down, and remaining flat		Symmetrical and skewed	
Steep and gradual		Wide and narrow	
Steady and fluctuating		Clusters and gaps	
Random and repeating		Tightly and loosely distributed	
Straight and curved		Normal and abnormal	

Beware of Patterns that Aren't Actually There

Humans' pattern recognition skills are amazing and the source of great insights, but sometimes they're too good. We are so adept at finding patterns that we sometimes detect ones that aren't really there. Even when a pattern is real, we often err by ceasing our exploration once we've spotted a single pattern, especially one that we were primed to find, so we miss others that are unfamiliar and unexpected. We should never become so good at spotting patterns that we lose sight of the information that composes them.

It is important, at times, to disregard familiar patterns and view data with fresh eyes. Take the time to look at the pieces—the details—for sometimes that is where truth lives. New patterns and meanings emerge that are unexpected if we let ourselves look without preconceptions and drill down to the specifics as well as scanning the big picture. Only when we empty our minds of the expected can we make room for something new. Zen Buddhism speaks of this approach as having a *beginner's mind*. In his marvelous book *Presentation Zen*, Garr Reynolds describes this state of mind:

Like a child, one who approaches life with a beginner's mind is fresh, enthusiastic, and open to the vast possibilities of ideas and solutions before them. A child does not know what is not possible and so is open to exploration, discovery, and experimentation. If you approach creative tasks with a beginner's mind, you can see things more clearly as they are, unburdened by your fixed view, habits, or what conventional wisdom say it is (or should be).⁵

Never let yourself become such an expert, so adept at spotting patterns, that you can no longer be surprised by the unexpected. Set the easy, obvious answers aside long enough to examine the details and see what might be there that you can't anticipate. Let yourself get to know the trees before mapping the forest.

5. *Presentation Zen*, Garr Reynolds, New Riders, Berkeley CA, 2008, p. 33