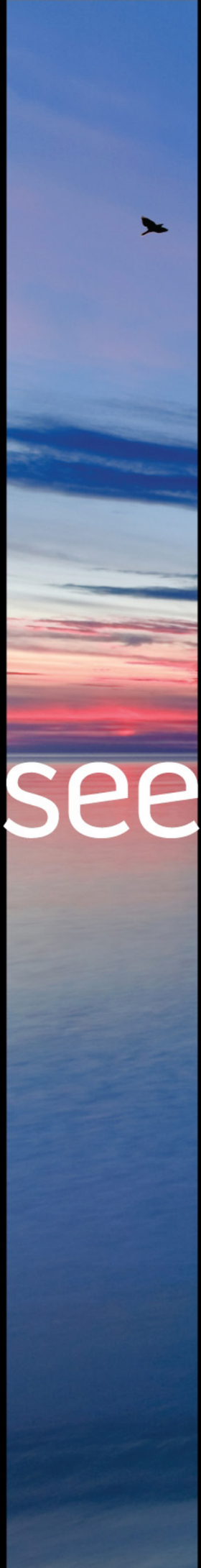
Simple Visualization Techniques
for Quantitative Analysis

Now you see it



Stephen Few

7 TIME-SERIES ANALYSIS

Introduction

No quantitative relationship receives more attention than values changing through time. "A random sample of 4,000 graphics from 15 of the world's newspapers published from 1974 to 1989 found that more than 75% of them featured time series."[1] Business analysts hope to see profits increase over time. Government and non-governmental organization (NGO) analysts expect to see changes in relation to influential events, such as natural disasters or tax increases. More than any other variable, time gives us a context for understanding data. The present can only be understood and the future can only be predicted in light of the past.

1. "An Augmented Visual Query Mechanism for Finding Patterns in Time Series Data," Eamonn Keogh, Harry Hochheiser, and Ben Shneiderman. *Proc. Fifth International Conference on Flexible Query Answering Systems*, Copenhagen Denmark, Oct 2002.

Time-Series Patterns

Six basic patterns are especially meaningful when we analyze change through time:

- Trend
- Variability
- Rate of change
- Co-variation
- Cycles
- Exceptions

Trend

A trend is the overall tendency of a series of values to increase, decrease, or remain relatively stable during a particular period of time. For example, it is common to refer to sales during a 12-month period as trending upward, downward, or remaining flat (also known as "no trend"). Any period of time can be the basis for determining a trend. We select a starting point and an ending point in time and then look to see whether the values during that period tended to move in a particular direction.

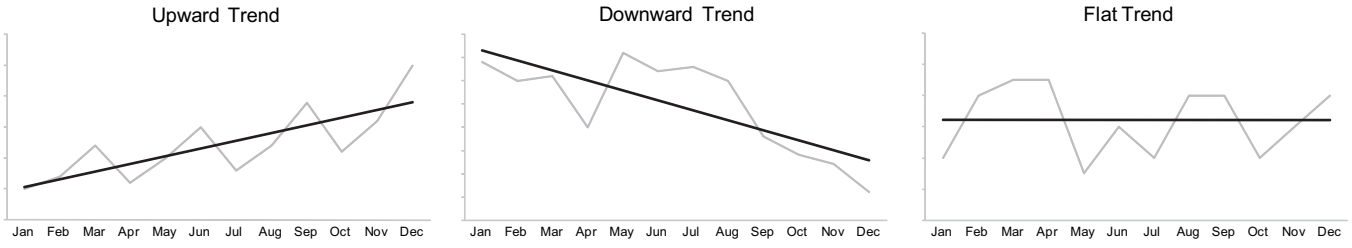


Figure 7.1

Line graphs work particularly well for visualizing trends. Trends are often obvious from the general slope of a line, but when the line moves both up and down throughout the period, the overall trend might be difficult to determine based on the appearance of the line alone (see the top graph below). At such times, most software can display a trend line to show the overall slope of change, but we must rely on trend lines with caution as I will explain later in this chapter, in the *Time-Series Analysis and Best Practices* section, where I propose an alternative to trend lines.

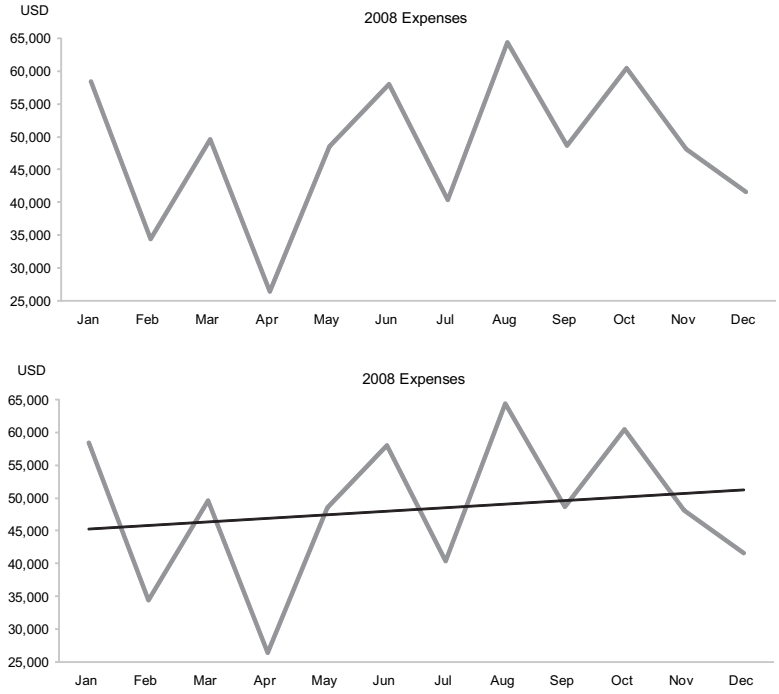


Figure 7.2

Variability

Variability is the average degree of change from one point in time to the next throughout a particular span of time. If sales revenues changed dramatically from month to month during a particular year, we can describe them as highly variable.

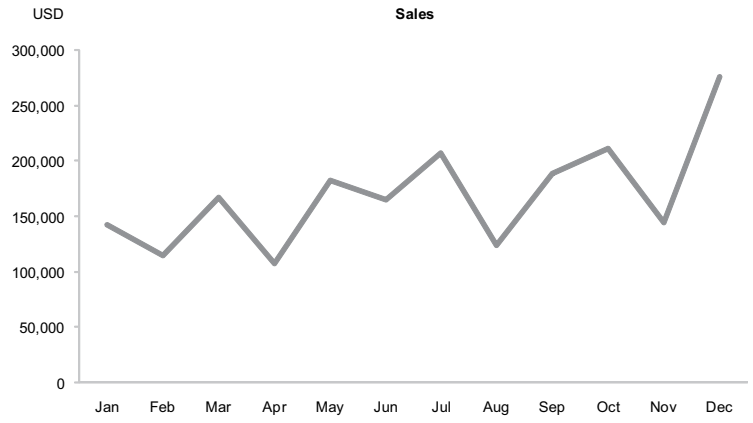


Figure 7.3

If sales decreased significantly from November to December, but remained fairly steady from month to month during the rest of the year, overall variability would not be considered high.

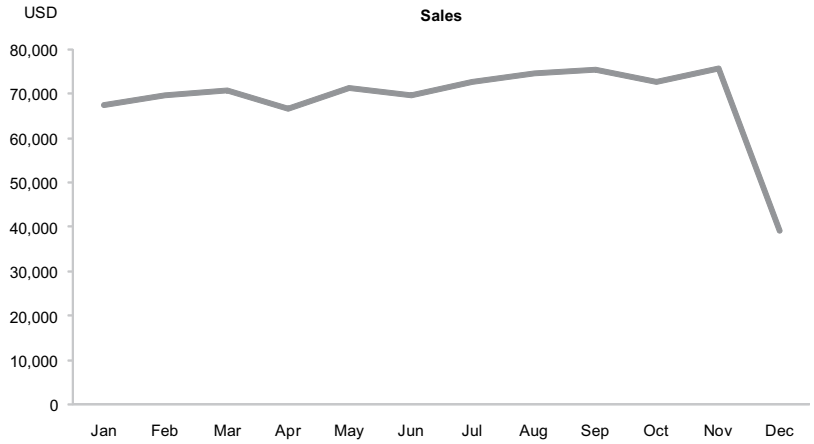


Figure 7.4

Line graphs do a good job of displaying variability. A jagged line indicates a greater degree of variability than another line in the same graph that is relatively smooth. Beware, however, of assuming a high degree of variability when viewing a graph that contains only a single jagged line. The appearance of jaggedness could be the result of a narrow quantitative scale. For example, in the graph below, revenues vary by less than 1% per month, but the pattern looks highly variable because its scale is so narrow.

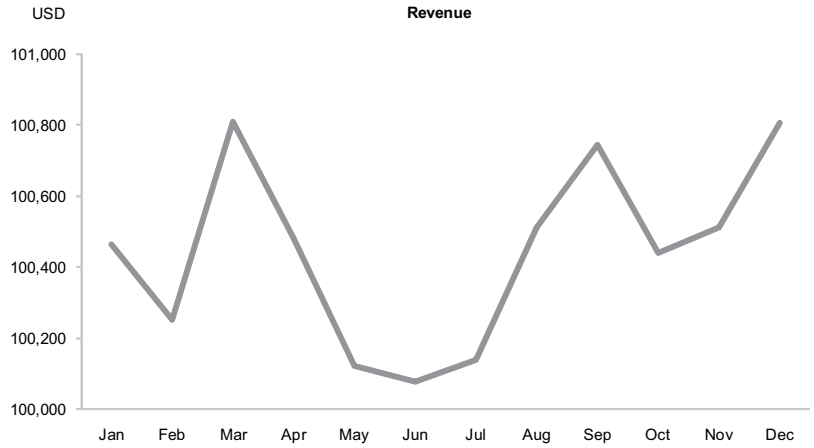


Figure 7.5

When interpreting variability based on the jaggedness or smoothness of a line, our judgments will be more reliable if we begin the quantitative scale at zero or, if examining a graph with multiple lines, restrict ourselves to relative assessments of more or less variability among the lines. In the following example, we can conclude with certainty that the monthly revenues in the east division experienced greater variability than those in the west division. We must be careful when judging the overall degree of variability in the east, however, because the narrow scale exaggerates its appearance.

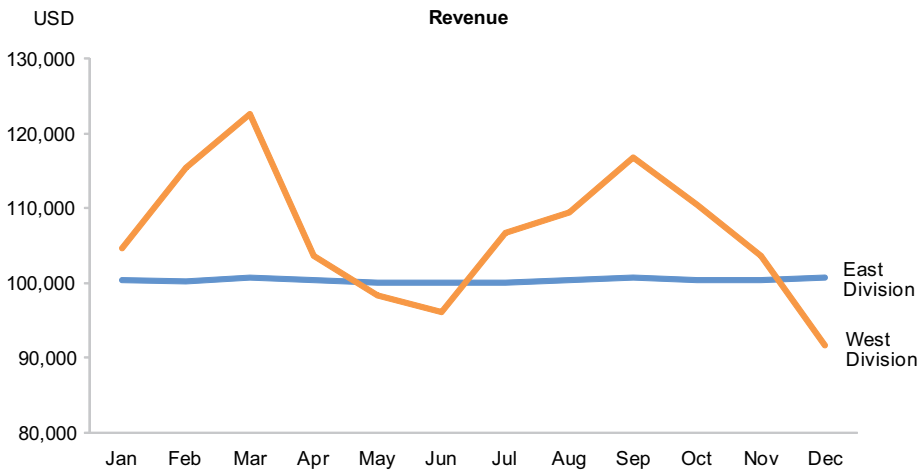


Figure 7.6

Rate of Change

The rate of change from one value to the next can be directly expressed as the percentage difference between the two. It is often enlightening to view change in this manner, especially when comparing multiple series of values, such as sales per region. For example, consider a comparison of domestic and foreign sales per month expressed in dollars, as illustrated below.

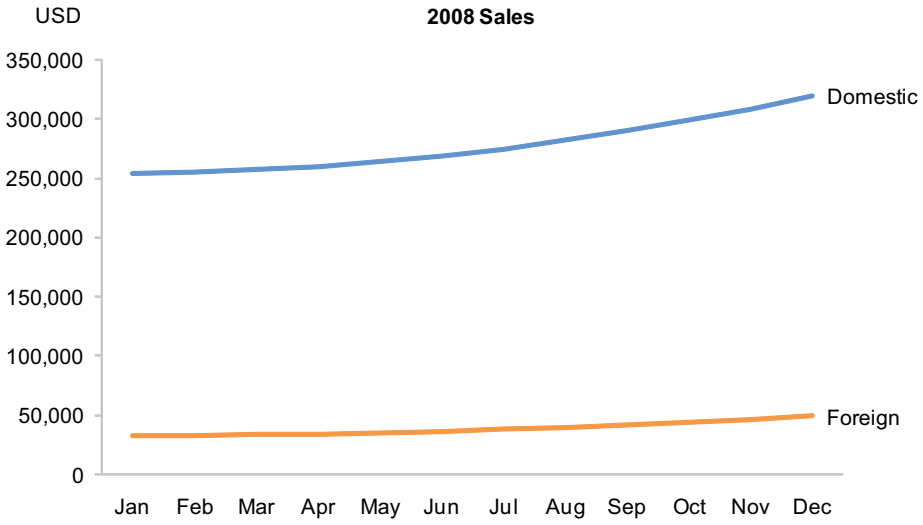


Figure 7.7

The amount of increase from month to month, measured in terms of U.S. dollars, is much greater for domestic sales than for foreign sales ($1,000 versus $250). However, this isn't a good way to compare rates of change between these two regions because even though foreign sales are increasing by smaller dollar

amounts, they might in fact be increasing at a faster rate. In the next example, we see the same sales data, this time expressed as the rate of change from one month to the next.

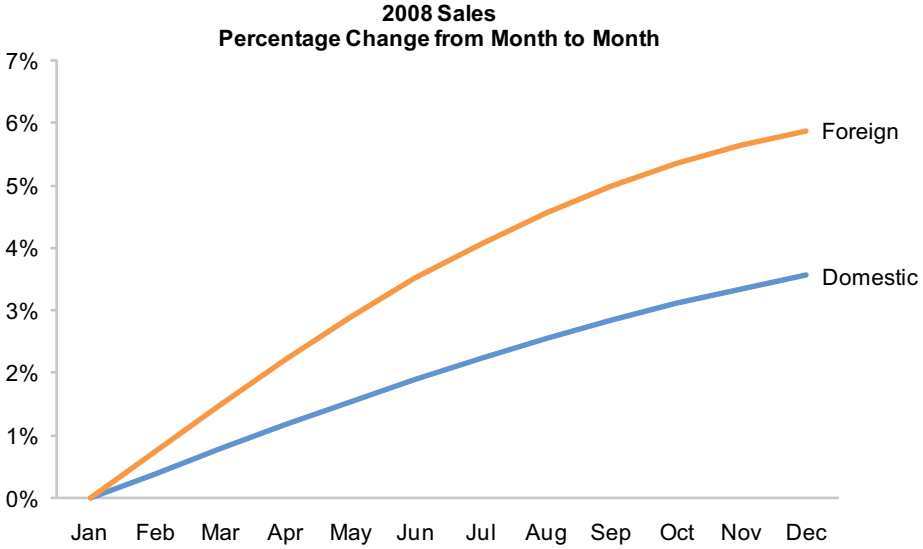


Figure 7.8

This graph tells a much different story than the previous one. Foreign sales are, in fact, increasing at a faster rate and thus might represent a better potential market. We'll look at various ways to examine rates of change later in the *Time-Series Analysis Techniques and Best Practices* section of this chapter.

Co-variation

When two time series relate to one another so that changes in one are reflected as changes in the other, either immediately or later, this is called co-variation. The pattern can qualify as co-variation even if changes in one time series move in a different direction (up or down) from corresponding changes in the other. For example, expenses could co-vary with profits such that decreases in expenses are reflected as increases in profits. When related changes don't occur simultaneously, but, instead, changes in one time series always occur before or after related changes in another, we have what are called *leading indicators* or *lagging indicators*. A leading indicator is a change that occurs in one time series that relates to a change that takes place in another at a later time. A lagging indicator is the reverse. The line graphs on the following page illustrate co-variation between newspaper ads (leading indicator) and orders (lagging indicator), which occur four days later.

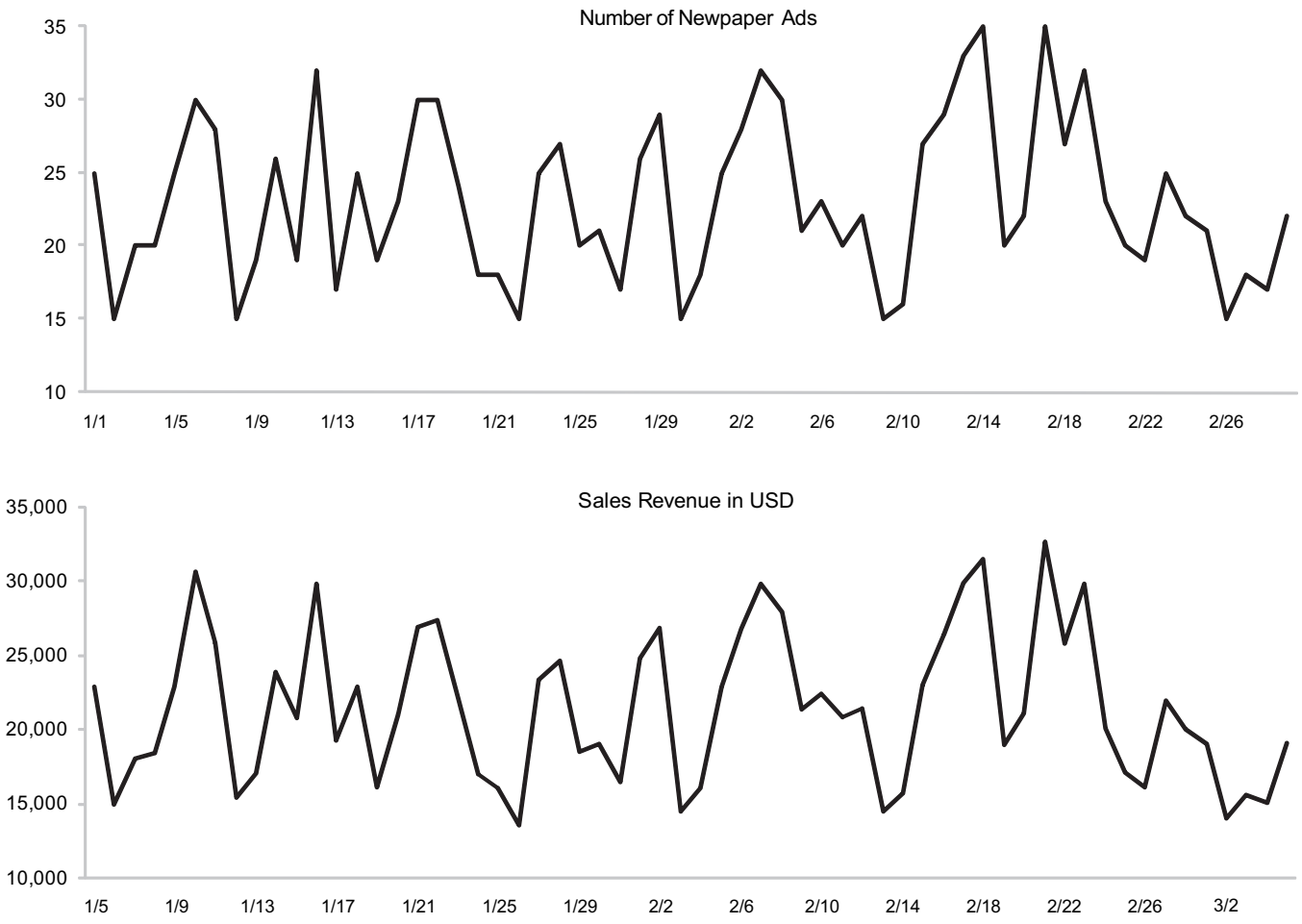


Figure 7.9

Cycles

All the patterns that we've covered so far are usually examined in a linear fashion, by viewing a period of time from beginning to end. For example, the question "At what time during the last five years did expenses hit their peak?" would be investigated using a linear view. Cycles, by contrast, are patterns that repeat at regular intervals, such as daily, weekly, monthly, quarterly, yearly, or seasonally (winter, spring, summer and fall). Cyclical patterns are often easier to examine using visualizations that don't display time linearly from beginning to end but instead display the interval at which the cycles occur (for example, days of the week, months of the year) positioned close to one another where they can be easily compared. The question "Did expenses consistently hit their peak during a particular month of the year during each of the last five years?" could be pursued using a cyclical view. The following line graph allows us to examine cyclical sales behavior by month in a manner that features quarterly patterns.

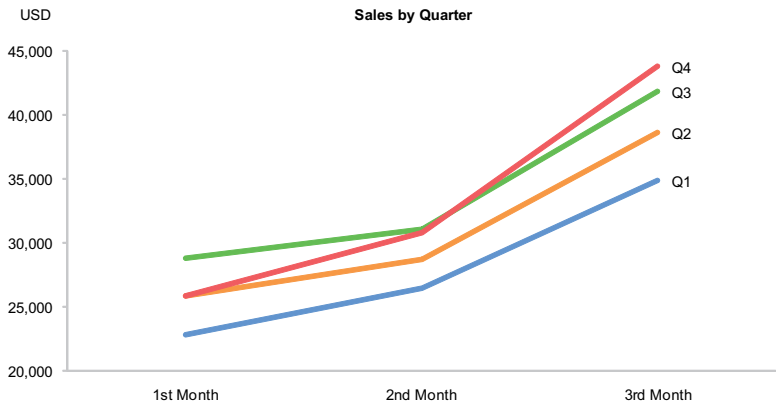


Figure 7.10

This particular sales pattern, which exhibits a peak in the last month of each quarter, is sometimes called the *hockey stick* pattern because it's shaped like a hockey stick with an upward bend near the end. If you've ever seen this pattern in your own company's sales data, you probably know that it is not the result of customer buying preferences but rather a result of the sales compensation plan, which awards bonuses to salespeople for reaching quarterly quotas. As the end of each quarter approaches, sales people get serious about closing as many deals as possible to reach or exceed their quota before the deadline. Once the end of the quarter is past, they relax for a while, sometimes on the golf course (sales-people do this, right?), until the next deadline looms.

Exceptions

We care about exceptions—values that fall outside the norm—in every type of analysis. How exceptions reveal themselves in graphs differs depending on the nature of the relationships that we're analyzing (time-series, distribution, correlation, and so on). In time series, they appear as values that are well above or below the norm, regardless of how we define the norm. In the following example, the number of employees hired during the month of November is a very visible exception, falling far below the number in other months.

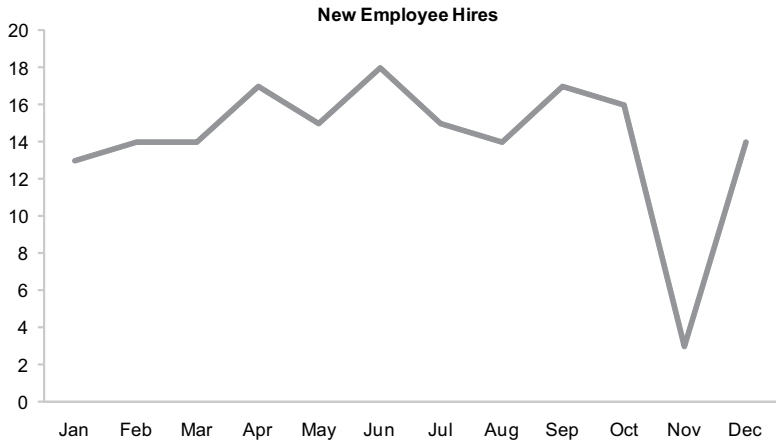


Figure 7.11

Every industry, work function, or group in an organization has particular and sometimes unique time-series patterns that are of special interest. Think about the data that you analyze and some of the patterns of change through time that are particularly interesting to your organization. Take a few minutes to list a few (or for "extra credit," you might even draw a few) of the patterns that are of interest in your work. Calling these patterns to mind when they are not in front of you helps to build them into memory in a way that makes them easier to spot.

Time-Series Displays

It's hard to beat a line graph for displaying change through time. Most time-series analysis can and should be accomplished using line graphs. Sometimes, however, other graphs do a better job. Five types of graphs are useful, some more than others, for examining quantitative change through time:

- Line graphs
- Bar graphs
- Dot plots
- Radar graphs
- Heatmaps

Each of these is the best solution for examining a particular type of time-series data or to help uncover a particular aspect of the data. Two more graphs are also useful for analyzing data when change through time is secondary in importance to another quantitative relationship:

- Box plots (and similarly constructed high-low plots), to analyze how distributions of values have changed
- Scatterplots, using animation to analyze correlation changes

Let's take a look at the design, uses, and benefits of each.

Line Graphs for Analyzing Patterns and Exceptions

If your objective is to see how quantitative values have changed during a continuous period of time, nothing works better than a line graph. Lines work better than any other means to make visible the sequential flow of values as they have changed with the passage of time. By its very nature, a line clearly traces the connection from one value to the next and through its slope displays the extent and direction of change. In the next example, the overall trend of sales throughout the year and the ups and downs from month to month are both easy to see in the line.

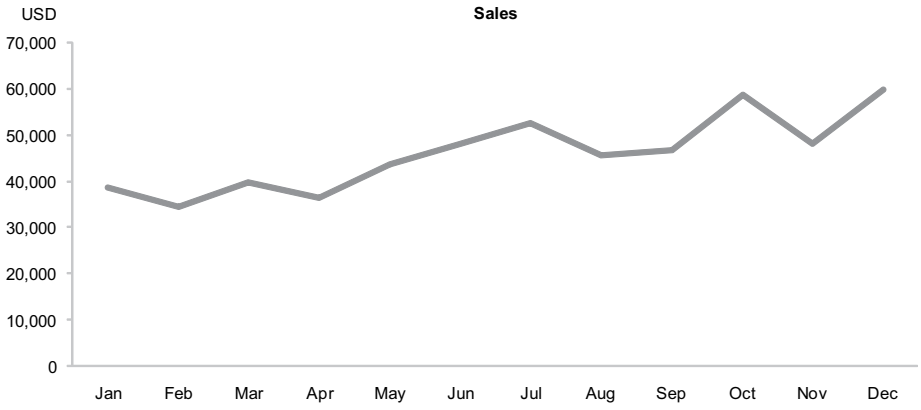


Figure 7.12

When we want to compare individual values at specific points in time, such as actual and budgeted expenses in a given month, bar graphs do this better. As I mentioned previously, if we wish primarily to see the shape of change through time, but to also compare the magnitudes of values at a particular point in time, we can include points along the lines to mark the precise location of each value, which makes it easier to compare values on separate lines at precisely the right location.

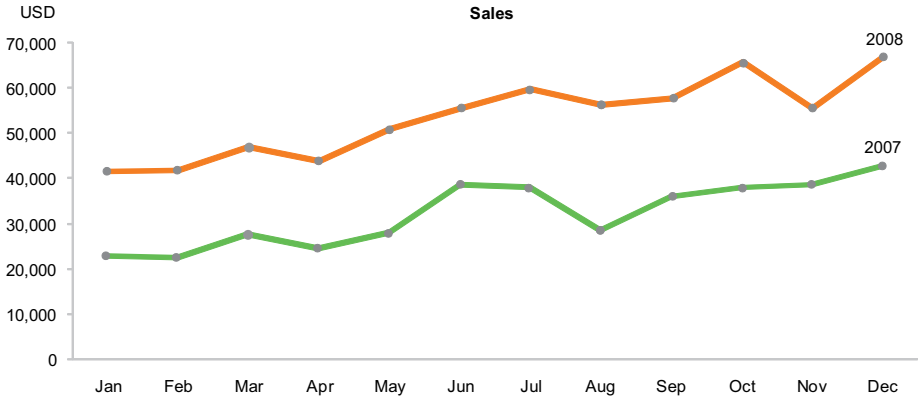


Figure 7.13

Bar Graphs for Emphasizing and Comparing Individual Values

As mentioned previously, the lengths of bars encode values in a way that allows us to simply and accurately compare individual values to one another. The visual weight of bars and their clear separation from one another encourages our eyes to focus on individual values rather than the patterns they form as a whole. Consequently, if you are trying to compare individual values, such as actual to budgeted expenses in particular months as illustrated on the next page, bars graphs do the job nicely.

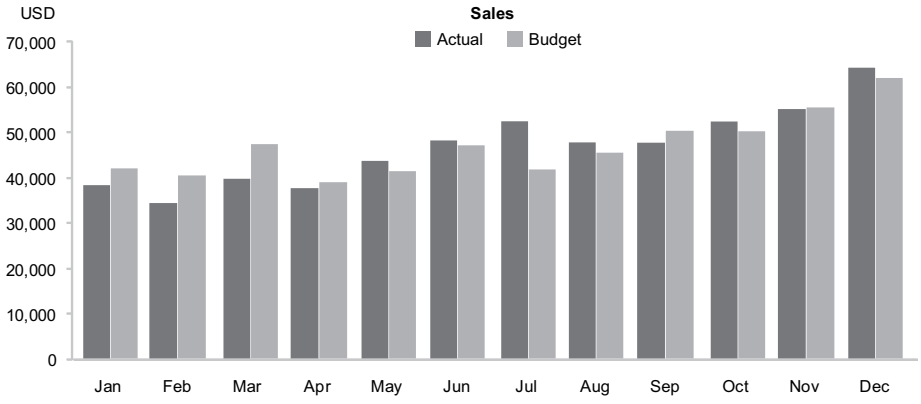


Figure 7.14

Dot Plots for Analyzing Irregular Intervals

Usually when we're analyzing time-series data, each interval along the timeline for each series of data contains a value. Sometimes, however, the values that we're examining are not distributed at regular intervals throughout the period but instead were recorded sporadically. For example, imagine that you're an inspector who measures contaminant levels at a particular location in a river, but not at regular intervals. If you used a line graph to display these values, it might look something like this:

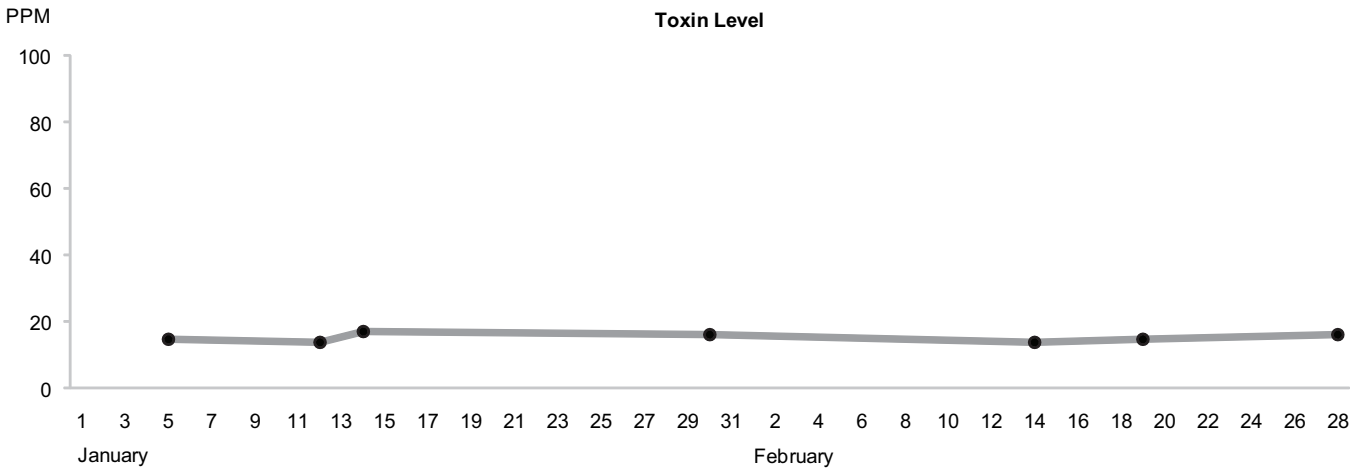


Figure 7.15

When we connect values that are located at irregular points in time with a line as I've done above, the resulting shape suggests a smooth linear change from one value to the next. This is a problem, however, because these smooth transitions might not at all correspond to what actually happened. If the toxin levels had been measured every day, the picture of change might look quite different, such as shown in the following graph.

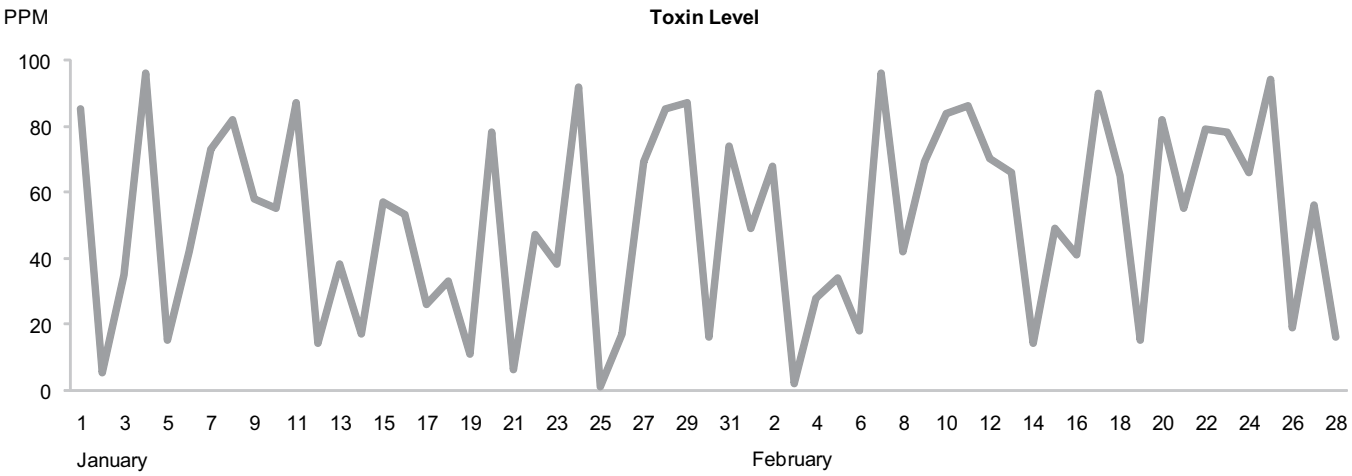


Figure 7.16

Therefore, when analyzing values that are spaced at irregular intervals of time, don't connect them with a line. Instead, use a data point, such as a dot, to mark each value separately. This type of graph is called a *dot plot*, illustrated below.

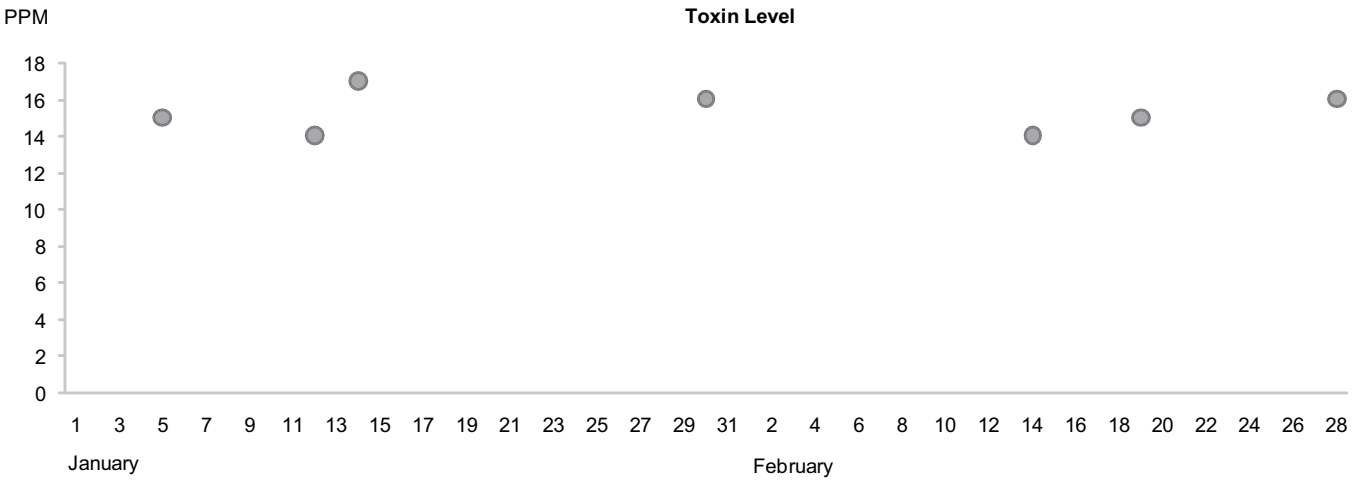


Figure 7.17

Dot plots discourage the misleading assumption that there was a direct transition from one value to the next. Few software products provide dot plots, but you can use many products to produce them, including Excel, by using a line graph with data points to mark the values and then eliminating the line.

Radar Graphs for Comparing Cycles

I'm not a big fan of *radar graphs* (also known as *spider graphs*) because their usefulness is limited to rare situations. Sometimes, however, they can be useful for time-series analysis. The circular shape of a radar graph can be used to represent the cyclical nature of time. For example, similar to the way hours are sequentially arranged in a circle on a clock, the axes of a radar graph can be used to mark the hours of the day, as shown in the example on the following page.

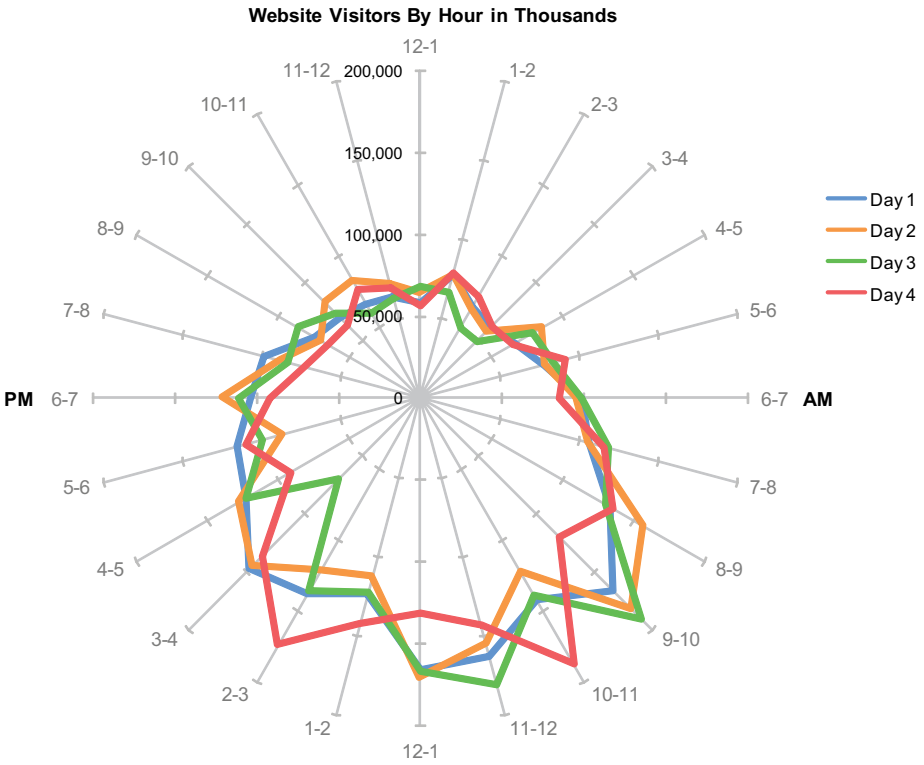


Figure 7.18

The same data can also be displayed using a line graph, as shown below, which I believe works just as well for analytical purposes. But if you prefer the way radar graphs represent the cyclical nature of time—the minutes of the hour, hours of the day, or even days of the week or month, months of the year, and so on—you'll find them useful.

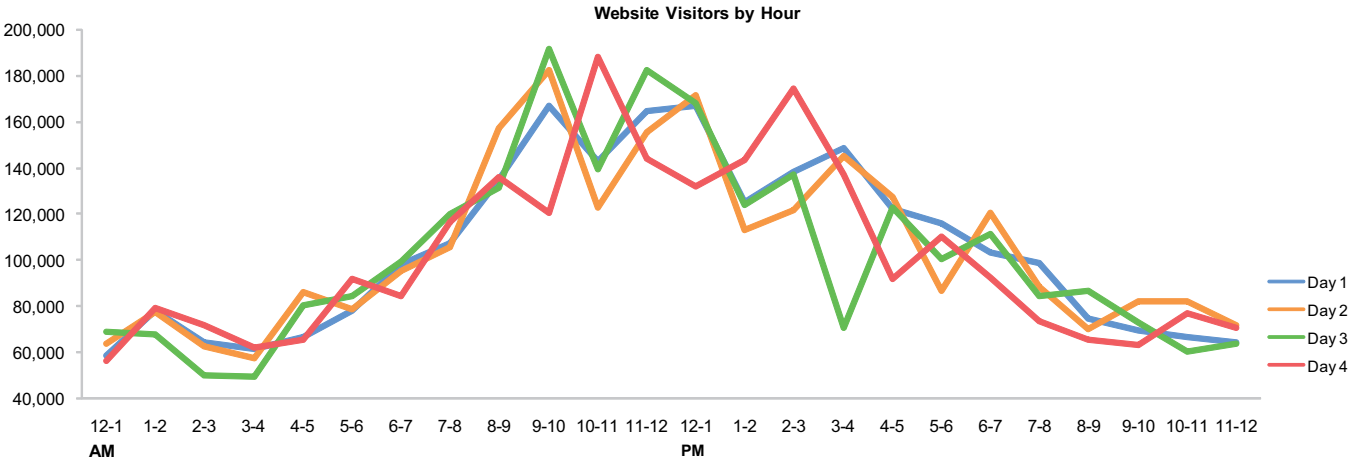


Figure 7.19

Both line and radar graphs can become cluttered when we use them to analyze cyclical patterns. The following example displays 30 days' worth of data, one line per day, resulting in a great deal of over-plotting, which makes it

impossible to compare individual days to one another or to closely examine anything that appears in one of the cluttered areas.

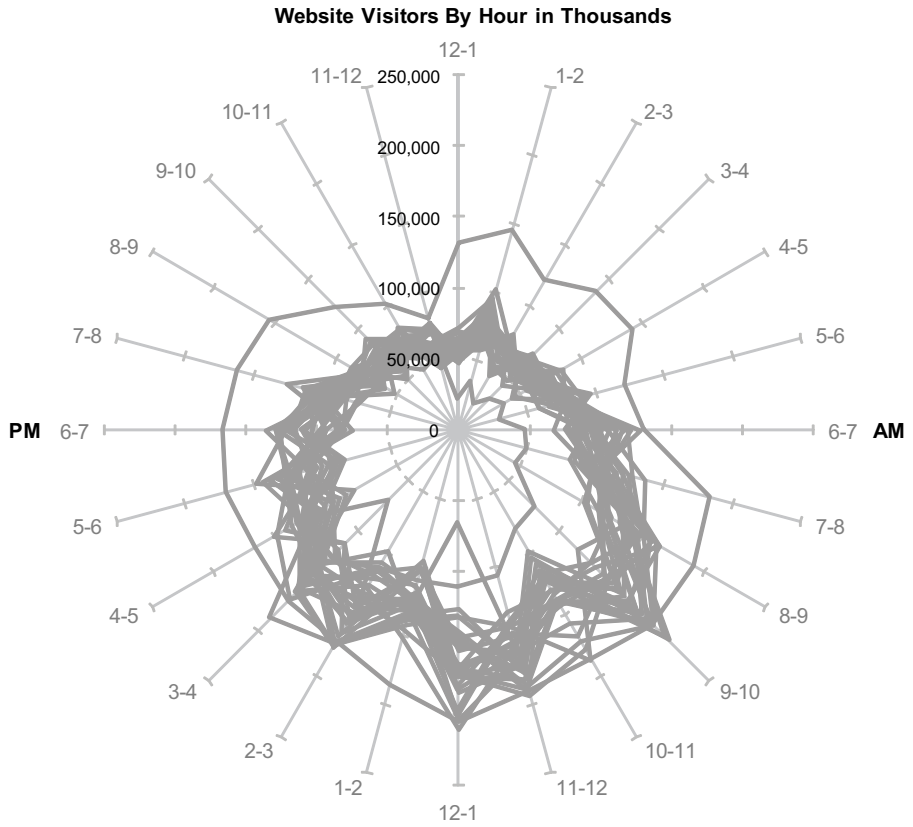


Figure 7.20

Despite the over-plotting, it is still possible to spot exceptions to the norm, such as the line that circles far outside the others or the one that is close to the center. It is also possible to discern predominant patterns, such as the high number of website visits that almost always occurs during the noon hour or the low number during the midnight hour. This is a useful overview of what's going on, which is a good place to begin. To dive down into the details using this display, we would need to reduce the over-plotting, such as by selectively filtering out days that aren't relevant to the question we're trying to answer.

Heatmaps for Analyzing High-Volume Cyclical Patterns and Exceptions

We can visualize large quantities of cyclical data that would likely produce over-plotting in a line or radar graph by using a *heatmap*. The term heatmap refers to any display that uses color to encode quantitative values. Weather maps are a familiar example of heatmaps. Typically, weather maps use variations in color on a geographical map to display temperatures or levels of rainfall. Another form of heatmap uses a matrix (rows and columns) of cells, each color coded to display a value. The following example, created using a product called

Trixie Tracker, is used by parents to track and attempt to understand the daily sleeping patterns of a young child over a period of one month.

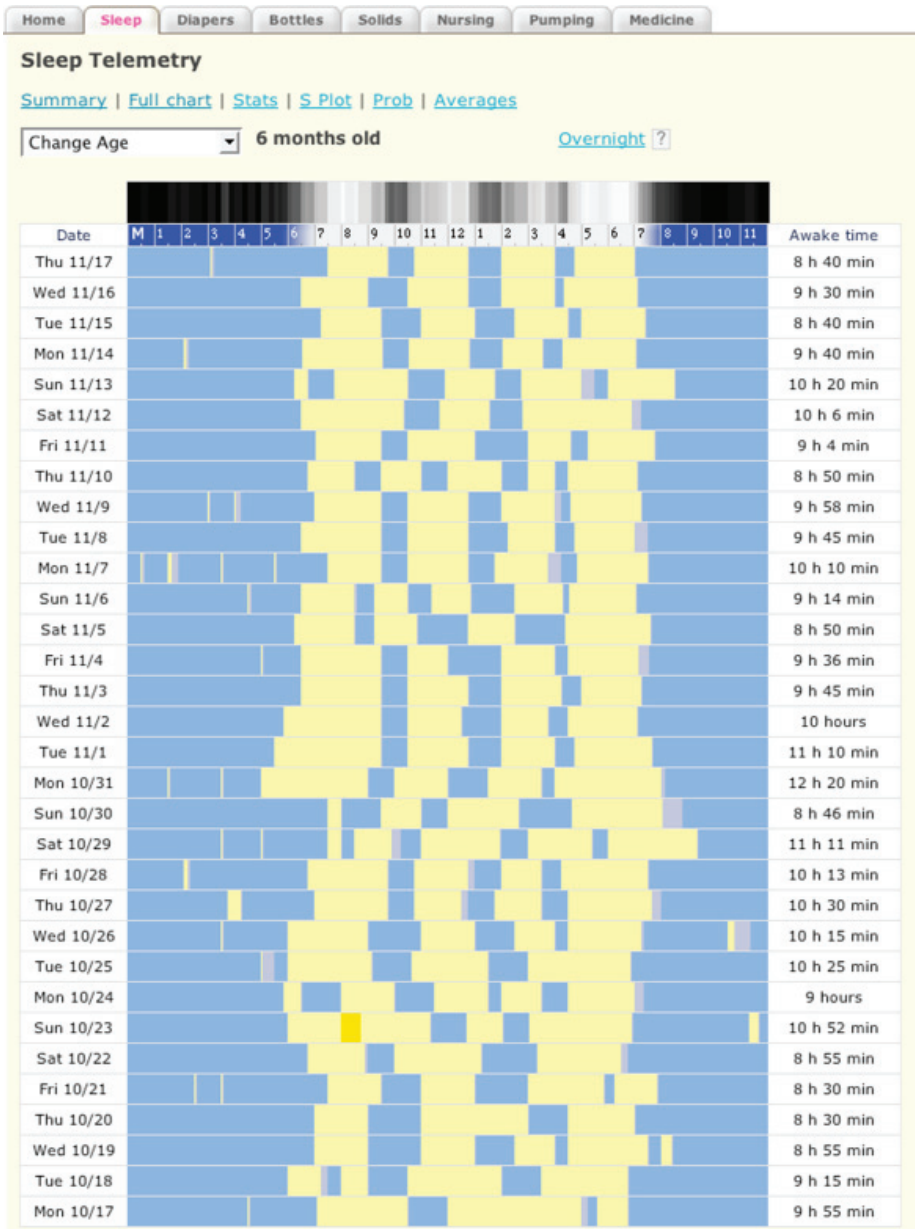


Figure 7.21. This chart was produced using a product called Trixie Tracker, which can be found at www.trixietracker.com.

Notice how easy it is to see the dominant patterns of awake time versus sleep time, especially using the summary in the row of grayscale colors at the top. This particular heatmap tracks and summarizes daily binary values (either on or off) of awake versus asleep, but heatmaps are not restricted to binary displays. The next example displays Web traffic, measured as the number of visits to a site during each hour of the day for 30 days. The number of visits in each hour has been encoded as varying intensities of red, with the highest values represented by the most intense color.

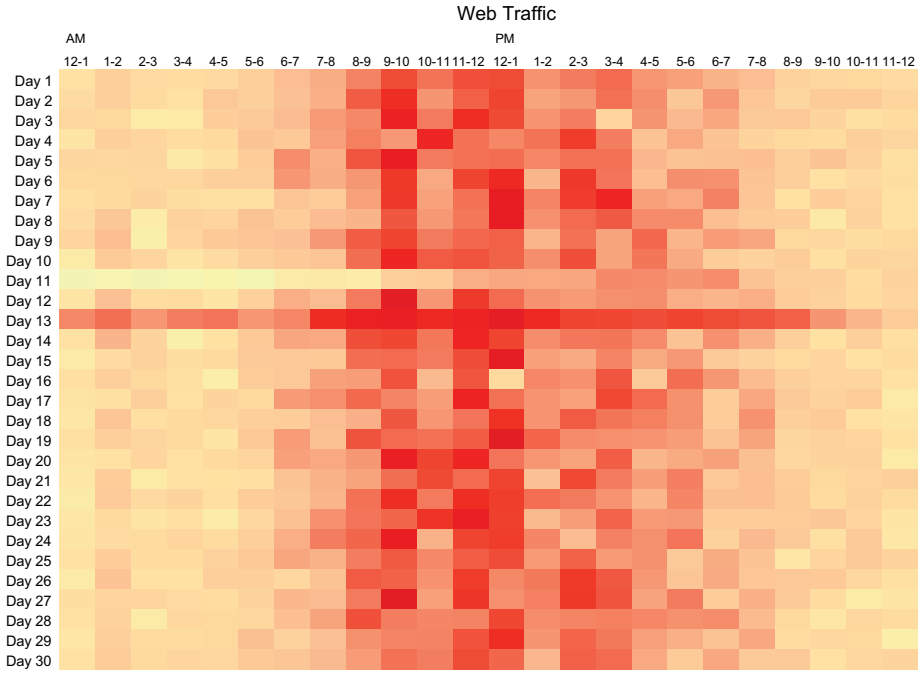


Figure 7.22

A heatmap matrix can be constructed in Excel using the *conditional formatting* feature. Other products handle heatmaps with more sophistication. Don't ever use a heatmap for time-series analysis just because it's novel or colorful. Use it only when it can display cyclical data that could not be as clearly displayed using a line or radar graph because of over-plotting.

Box Plots for Analyzing Distribution Changes

Box plots work superbly for analyzing how values are distributed across a range and how that distribution changes through time. For example, imagine that you're an analyst working in the Human Resources department of a large company; you've been asked to examine how salaries are distributed across the full range from the lowest to highest and how that distribution has changed during the past five years. A box plot such as the one below could be used for this task.

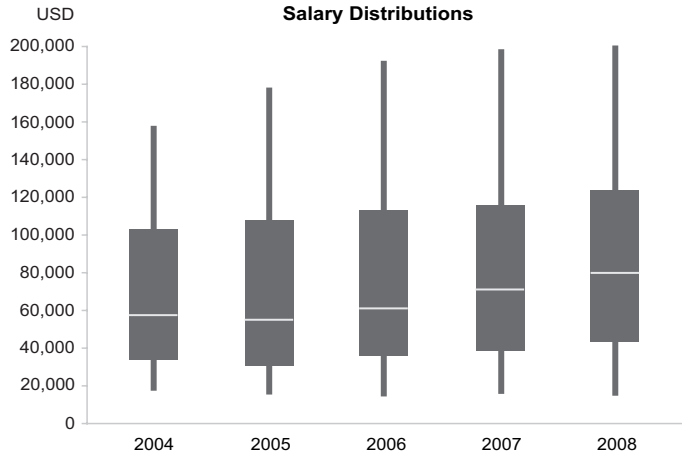


Figure 7.23

If you're not already familiar with box plots (most people aren't), don't worry. We'll spend quite a bit of time on them in *Chapter 10: Distribution Analysis*, and you'll become comfortable with them in no time at all. For now, here's an abbreviated version of the story that's told by the previous example. The typical salary paid in 2005 of about $56,000 (the light horizontal line that divides the box near the middle, which represents the median salary) was slightly lower than it was in 2004 (about $58,000), as was the lowest salary (the bottom of the vertical line). The highest salary (the top of the vertical line), however, increased significantly from around $158,000 to $179,000.The bottom half of salaries were crowded into a fairly narrow $41,000 range, compared to the top half, which were more liberally spread across a $123,000 range. In 2006, the typical salary switched directions and increased a fair amount, as did the highest salary, which happened again in 2007. In the final year, 2008, although 50% of the employees made less than $80,000 (far lower than the midpoint between the highest and lowest salaries near $100,000), salaries were more evenly distributed across the range than they ever were previously during this five-year period.

This next example is the same as the previous, except that the median values have been connected from one point in time to the next with a line to make it a little easier to see how the salaries have changed through time. This version of a box plot is not available in most products, but I find it quite useful for displaying how distributions have changed through time.

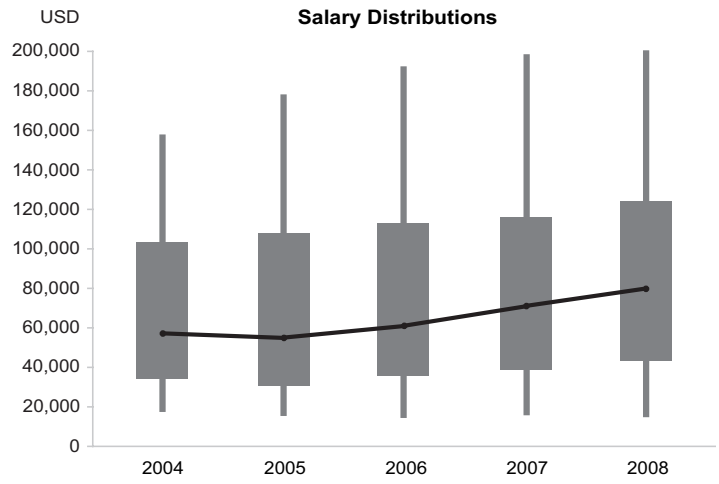


Figure 7.24

Animated Scatterplots for Analyzing Correlation Changes

Scatterplots are a wonderful way to compare two quantitative variables to determine whether, how, and to what extent they are correlated. We'll talk a lot about scatterplots in *Chapter 11: Correlation Analysis*; for now, I only want to describe how data points can be animated (given motion) to display how the relationship between two sets of quantitative values changed through time. This is accomplished by moving the data points around in the scatterplot to show

how values changed from one point in time to the next. This technique was pioneered and has been popularized by Hans Rosling of GapMinder (www.GapMinder.org), a Swedish professor and social scientist who uses it to tell important statistical stories. Here's an example that Rosling created, which shows this relationship between fertility rates and child mortality by country, grouped into continents by color, as it existed in 1962.

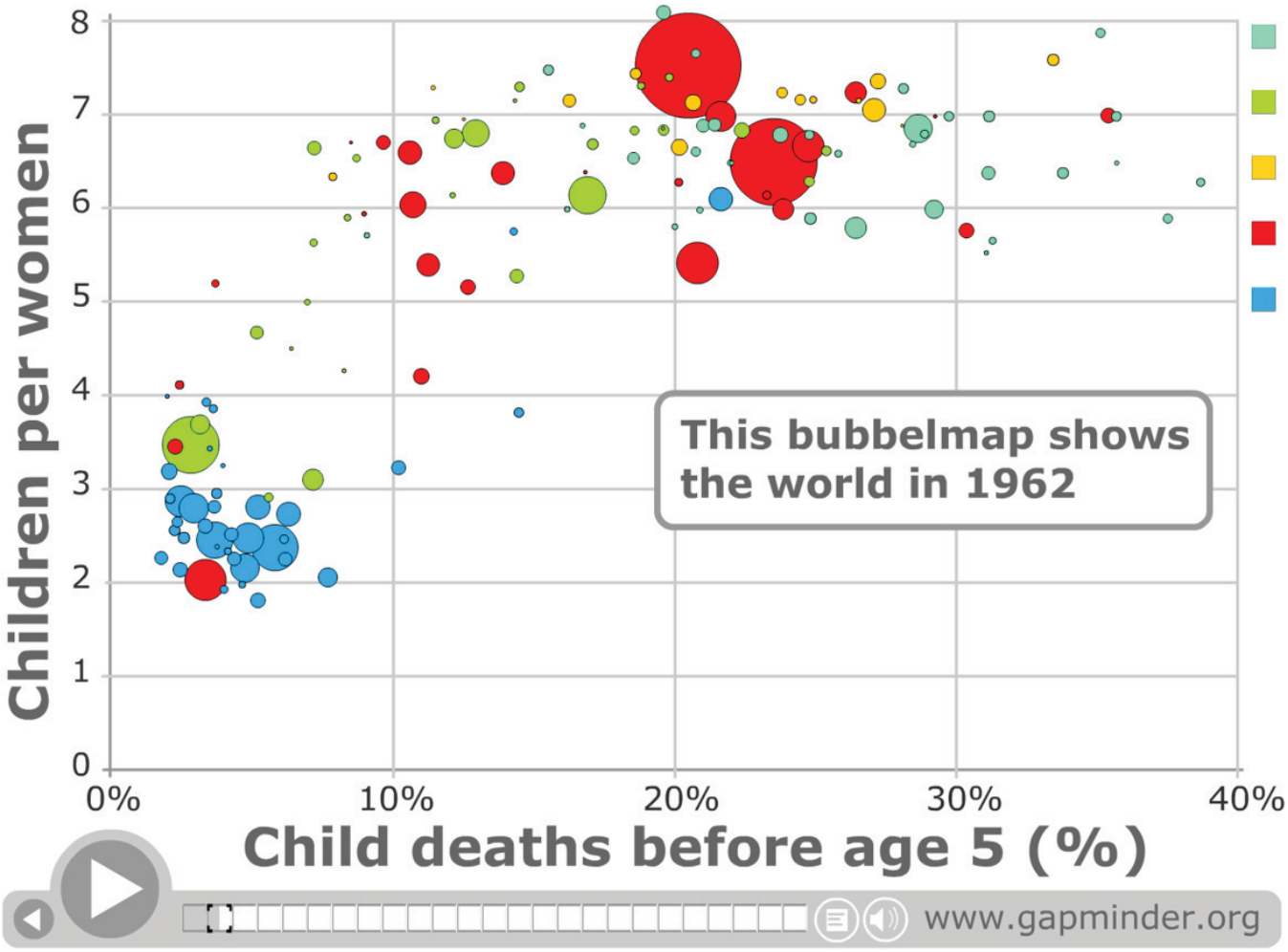


Figure 7.25. This and many other wonderful examples can be viewed in animated form at www.GapMinder.org. The software that was originally developed by Gapminder, called *Trendalyzer*, was purchased by Google, and a new version called *Motion Charts* is now freely available.

When Rosling appeared for the first time at the Technology, Entertainment, and Design (TED) conference in Monterey, California in 2006, he had 20 minutes to tell the story of world fertility rates (number of births per woman) related to life expectancy (years). He stood in front of a huge projection of a chart while bubbles (one per country) moved around the screen to show how these values of fertility and life expectancy have changed from 1962 to 2003. People sat in awe as he ran around pointing to bubbles and speaking until he was winded, mesmerized by the story and his highly animated presentation style. Perhaps never before in history did a crowd of people find a bubble chart

so fascinating. Why? I believe both because the story itself was compelling and important and because the animated bubble chart brought the story to life in a way that made it easy to understand.

Figure 7.26. Hans Rosling during his presentation at the TED conference in 2006.

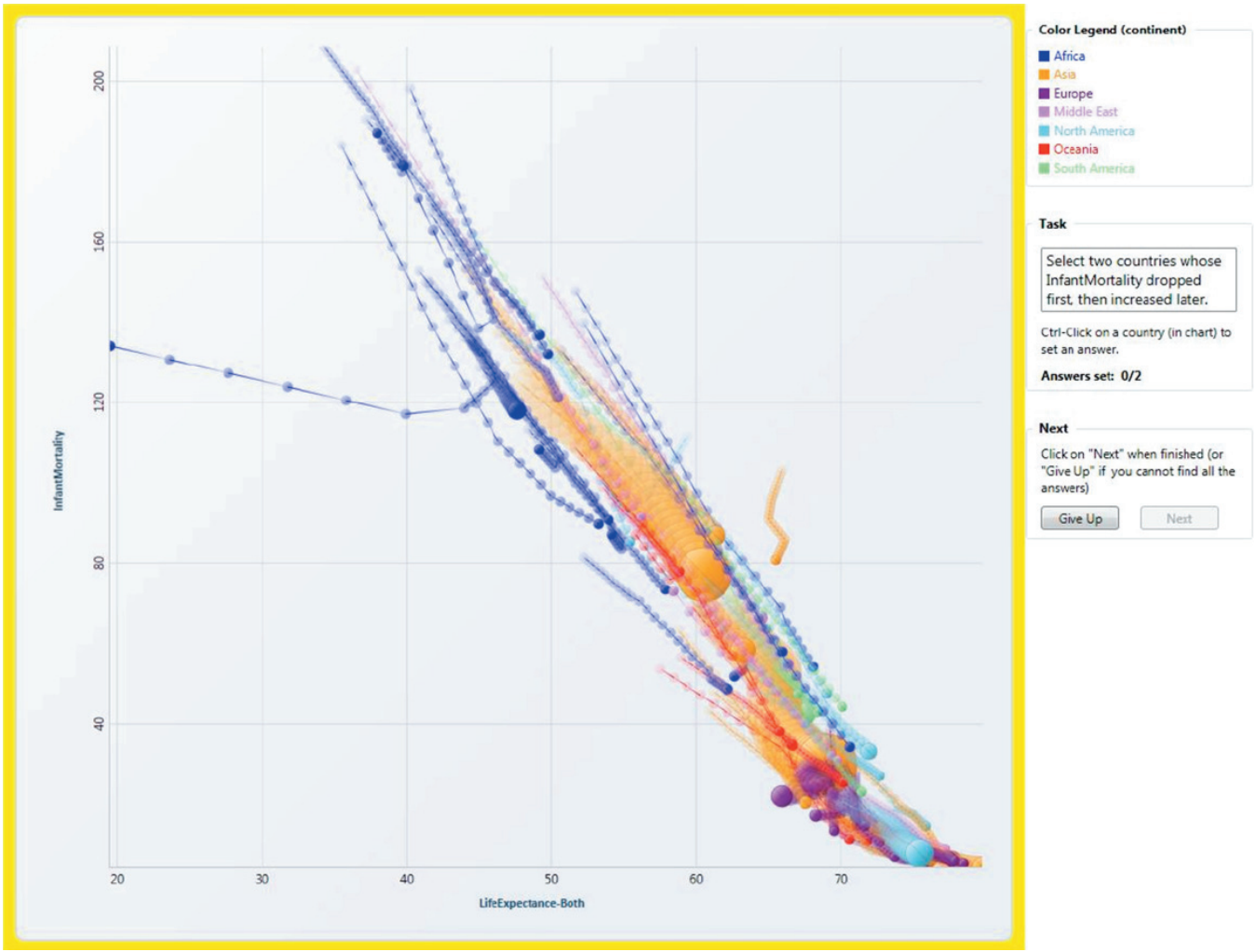
Animations can be used in powerful ways to tell the story of change through time. Of this I have no doubt, but, for our purposes, the question is: "Can animations of change through time be used for analysis?" Several researchers recently tackled this question, conducting a series of experiments with enlightening findings. Animation works very effectively for telling a story because a narrator tells us where to focus our attention as facts unfold across the screen. It does not demonstrate the same benefits when used for analysis. If we're trying to watch all the little bubbles as they move around, we can only take in a fraction of what's going on. To make sense of it, we end up rerunning the animation over and over, attending to a different bubble or two each time, which is not only time consuming, it also makes it impossible to stitch the pieces together into a big picture of what went on because, as you recall, our working memory is limited.

"Effectiveness of Animation in Trend Visualization," George Robertson, Roland Fernandez, Danyel Fisher, Bongshin Lee, and John Stasko, *IEEE Transactions on Visualization and Computer Graphics*, Volume 14, Number 6, November/December 2008.

For analytical purposes, times-series animations must be supplemented by other displays that allow us to follow what happened, discern the pattern of change, and make comparisons. The study of animation for data analysis confirmed the effectiveness of two approaches that allow us to perform these tasks:

- Trails to show the complete pattern of change through time from start to finish
- Small multiples, such as trellis displays, to compare the patterns of change among multiple items

Rosling uses trails in the form of a separate bubble per interval of time (for example, for each year) to help people see and compare patterns for the entire span of time. This study of animation techniques improved the effectiveness of trails by connecting each bubble that represents a point in time with a line and using color intensity from light to dark to show the direction of change along the trail.



By looking at a display like this, we can discern predominant patterns and outliers, but we cannot easily compare specific patterns because of the visual clutter and over-plotting. Trellis displays solve this problem by separating each of the trails into its own graph, as you can see in the next example.

Figure 7.27. "Effectiveness of Animation in Trend Visualization," George Robertson, Roland Fernandez, Danyel Fisher, Bongshin Lee, and John Stasko, *IEEE Transactions on Visualization and Computer Graphics*, Volume 14, Number 6, November/December 2008, p. 1327.

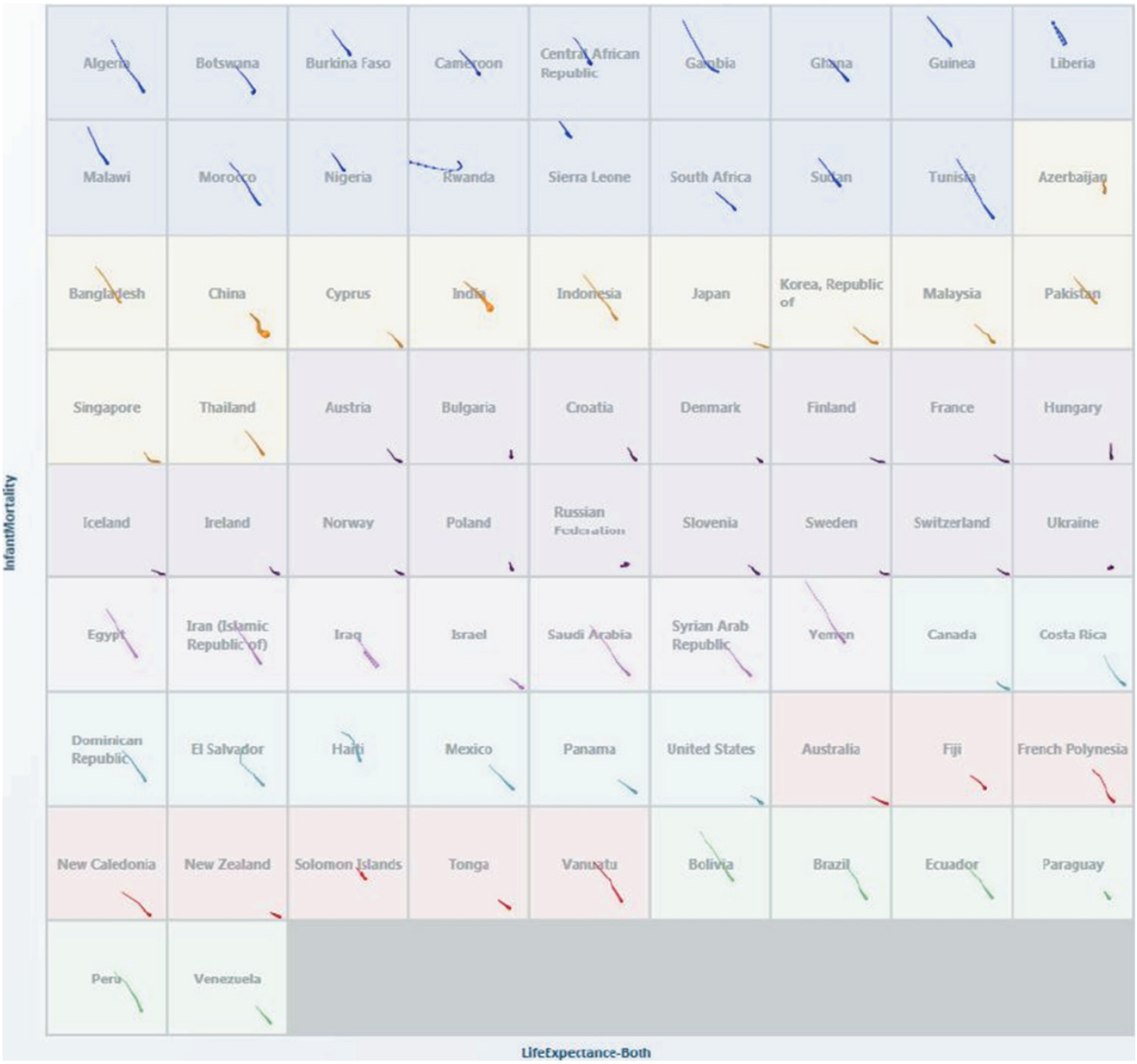


Figure 7.28. "Effectiveness of Animation in Trend Visualization," George Robertson, Roland Fernandez, Danyel Fisher, Bongshin Lee, and John Stasko, *IEEE Transactions on Visualization and Computer Graphics*, Volume 14, Number 6, November/December 2008, p. 1328.

Thanks to the innovative efforts of Rosling and fine-tuning by several exceptional researchers, we now know the usefulness of time-series animations, their limitations for analysis, and alternative complementary displays that enable us to see and compare patterns of change.

Time-Series Analysis Techniques and Best Practices

Given how important time-series analysis is to most organizations, it is no surprise that several techniques have been developed to peak under the covers to observe time's mysteries. It should also come as no surprise that a few guidelines (best practices) should be followed to avoid mistakes in visualizing and analyzing time series.

We'll look at the following techniques and best practices:

- Aggregating to various time intervals
- Viewing time periods in context
- Grouping related time intervals
- Using running averages to enhance perception of high-level patterns
- Omitting missing values from a display
- Optimizing a graph's aspect ratio
- Using logarithmic scales to compare rates of change
- Overlapping time scales to compare cyclical patterns
- Using cycle plots to examine trends and cycles together
- Combining individual and cumulative values to compare actuals to a target
- Shifting time to compare leading and lagging indicators
- Stacking line graphs to compare multiple variables
- Expressing time as 0-100% to compare asynchronous processes

Aggregating to Various Time Intervals

Have you ever noticed that time-series data can look quite different if you change the level of aggregation? For example, if you're examining a year's worth of visits to your website per quarter, and then per month, and then per day, the patterns of change revealed at each level might look quite different.

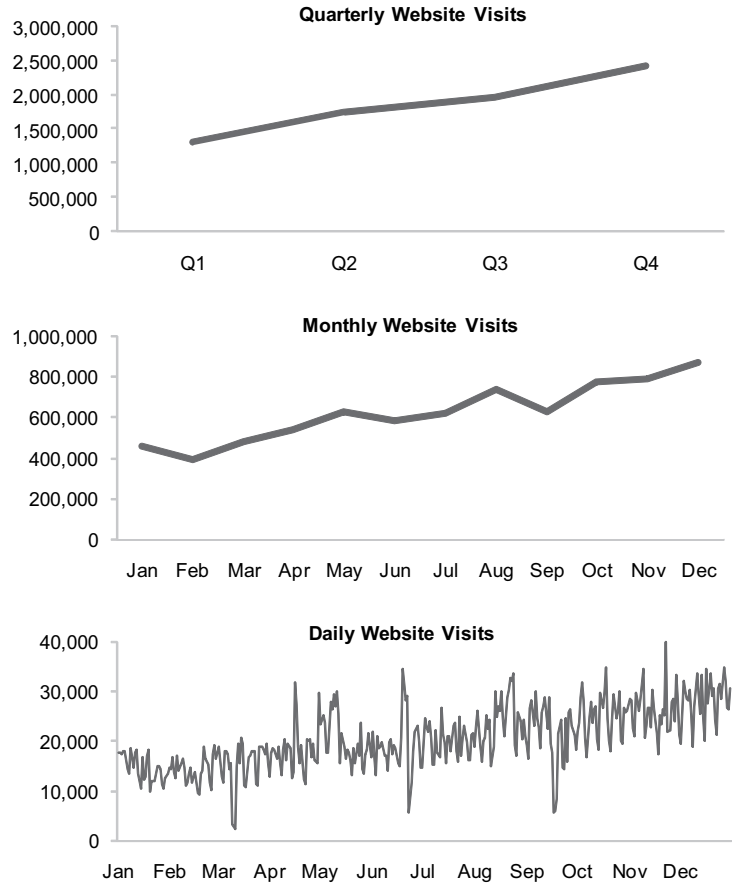


Figure 7.29

All three versions of the previous graph are useful and correct. The daily version reveals details that aren't visible when viewing the same data by month or quarter, such as the fact that Web traffic is higher on weekdays than on weekends. However, the overall trend is difficult to discern from the daily view. One view isn't better than the others in general, but one is definitely better than the others for specific analytical purposes. Don't restrict your view of time series to a single level of aggregation, especially when searching without preconceptions for anything that seems interesting. Switch the level from year to quarter, quarter to month, month to week, week to day, and so on (and back and forth) to tease out the insights that will only emerge when we look at the data from all perspectives.

To encourage this practice, software tools are needed that allow us to quickly and easily switch between various intervals of time while examining data. The ability to switch time intervals with a mouse click or two, or by using something as simple as an interval slider control, sets us free to explore.

Viewing Time Periods in Context

It is easy to read too much into a time-series pattern when viewing a brief span of time. This can easily happen if we restrict our analysis to a relatively short period and never view it in the context of a longer period. Consider the following example.

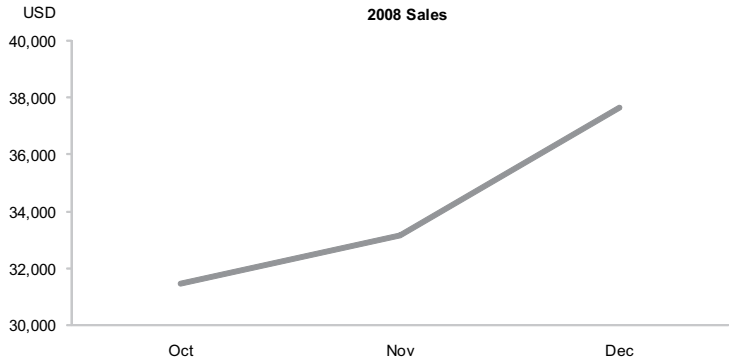


Figure 7.30

Based on this graph, we might conclude that sales are trending upward. Now look at the same three months of data, this time in the context of the entire year.

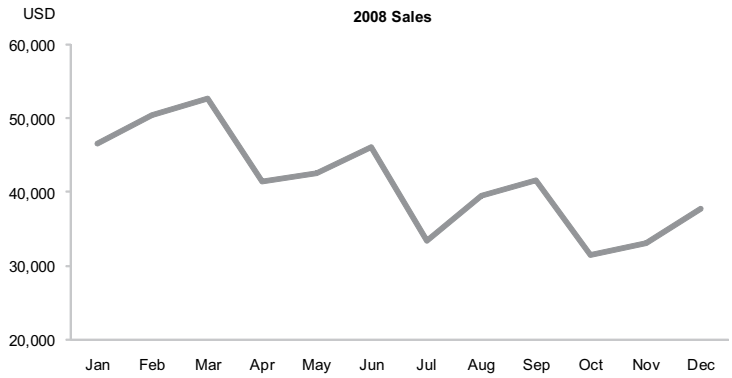


Figure 7.31

The lesson is clear, isn't it? When we examine short periods of time in isolation, we run the risk of assuming that observed patterns are more significant or more representative of what's happening overall than they in fact are. Is a year's worth of data enough? Five years? Ten years? There is no single right answer. Develop the habit of occasionally extending your view to longer stretches of time. Views of various time spans might each lead to insights that are not available if we stick to one time span.

Grouping Related Time Intervals

Sometimes it's useful to arrange time intervals into larger groups. For example, if we're examining three years' worth of monthly expenses, it would be useful to see the months grouped by year and perhaps also by quarter. As you can see in the second graph below, the addition of vertical lines to divide the quarters makes it easier to examine and understand quarterly patterns.

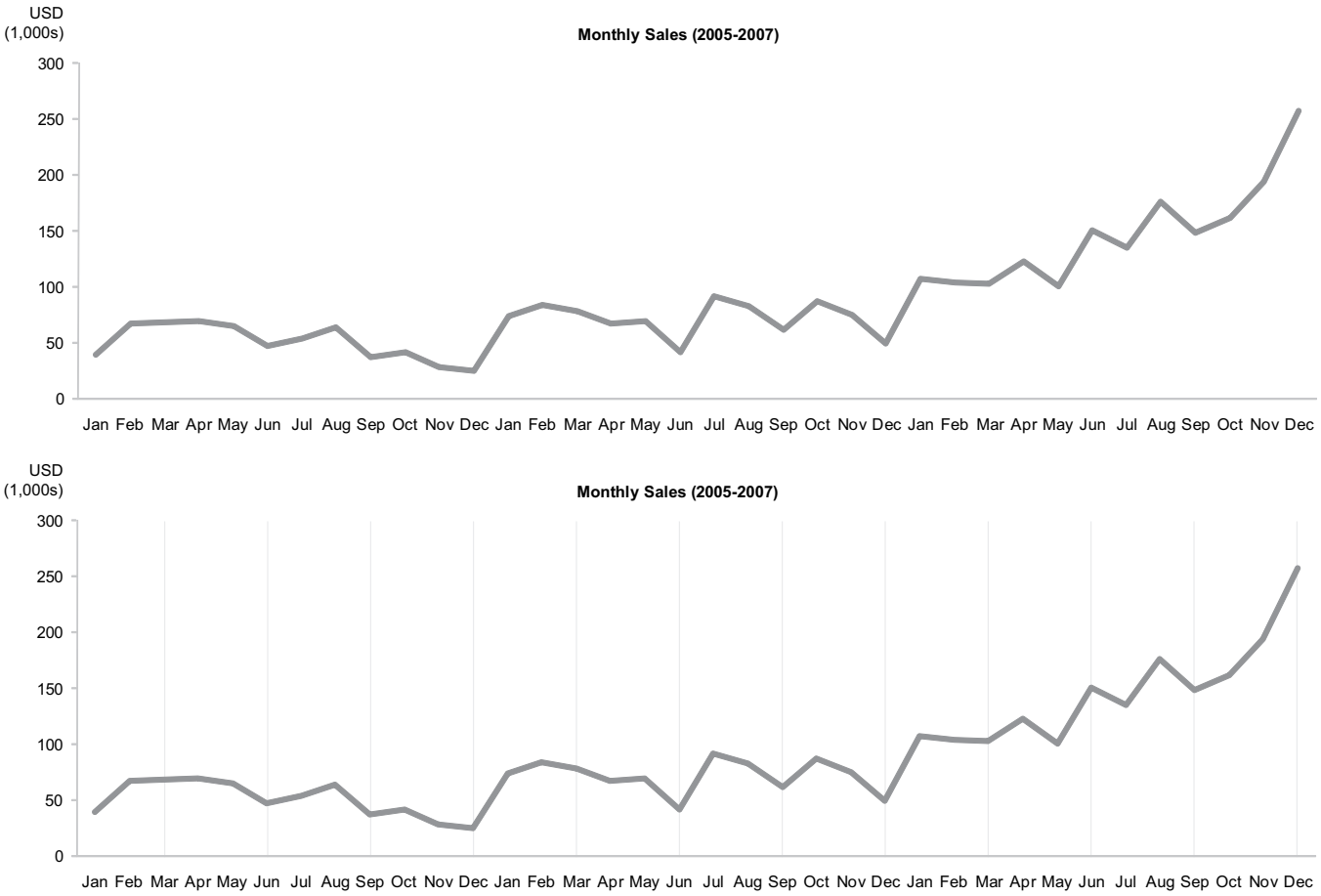


Figure 7.32

Another useful example is illustrated on the next page. When we are viewing three months' worth of daily website visits, it helps to clearly separate weekdays from weekends so we can, for example, differentiate expected drops in Web traffic on weekends from unexpected drops on weekdays, which ought to be investigated.

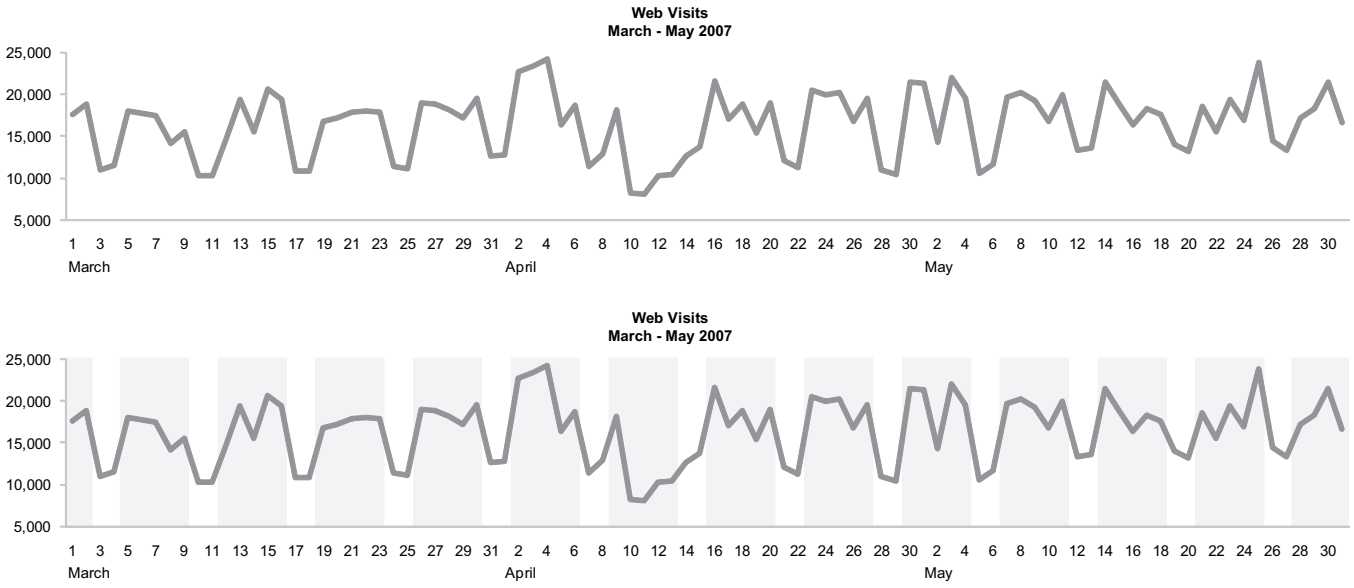


Figure 7.33

Unfortunately, few software products support the ability to group periods of time in this manner. If yours does not, let your vendor know how useful this would be.

Using Running Averages to Enhance Perception of High-Level Patterns

It is sometimes difficult to discern the overall trend of values during a span of time, especially when values are highly volatile. It's hard to picture how the general pattern might look if we could smooth out a jagged line of time-series values, taking all the increases and decreases into account, to discern what's happening overall. Trend lines are often used in attempts to solve this problem, but this approach can be misleading.

Bear in mind that whenever you take advantage of a software product's generous offer to draw a trend line for you, you are not only trusting it to do so accurately, you are also asking it to display a trend across a particular stretch of time that will fail to account for what has happened before or after. In the following example, I allowed Excel to display the overall trend of expenses for the current year to date (January through November).

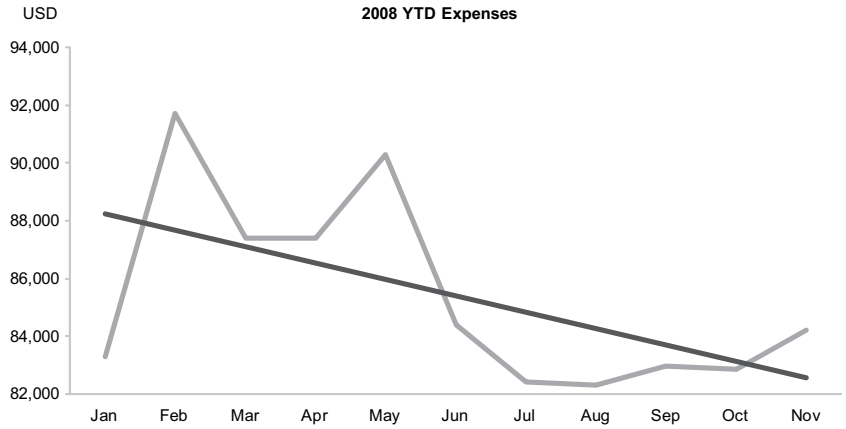


Figure 7.34

As you can see, the black trend line suggests that expenses are trending downward. Now look at how different the trend looks when I add a single month—December from the previous year—to include a full 12 months of expenses:

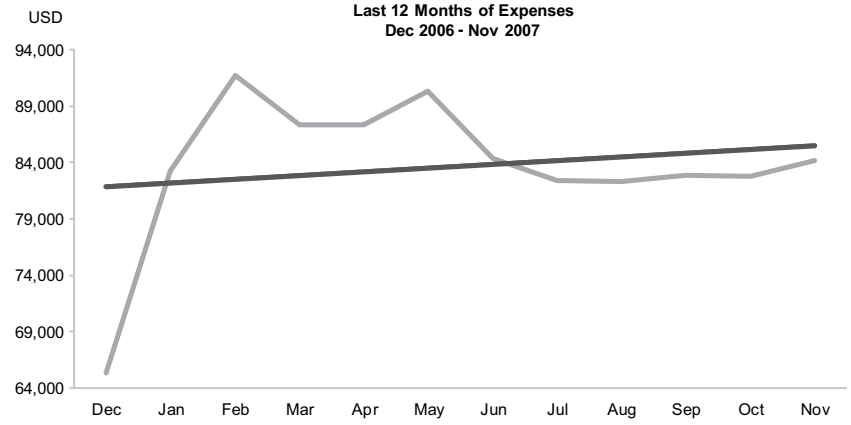


Figure 7.35

Quite a different trend, isn't it? Both graphs are accurate, based on the data they were asked to include when calculating the trend. If you associate a trend line with time-series data, be sure to examine values that fall outside the specified time period you're basing it on, to make sure you haven't isolated a section that would trend quite differently if the period were slightly altered.

A straight line of best fit, which is the type of trend line that appears in both examples above, is based on a calculation called a *linear regression*. It's determined by finding the straight line that passes through the full set of values from left to right such that the sum of the squares of the distance between each data point and the trend line is the least possible. Unless you understand this calculation and its proper use when applied to time series, it's easy to get into trouble. For this reason, I suggest a different approach to solving the problem: *running averages*.

Variability in time series can be smoothed out to some degree if, rather than displaying the actual value for each point in time (for example, for each month in the graph below), we display an average for each value and a few that precede it. For example, in the graph below we can see the pattern formed by taking the same values that appear in the two examples above and displaying each month's value as a five-month running average.

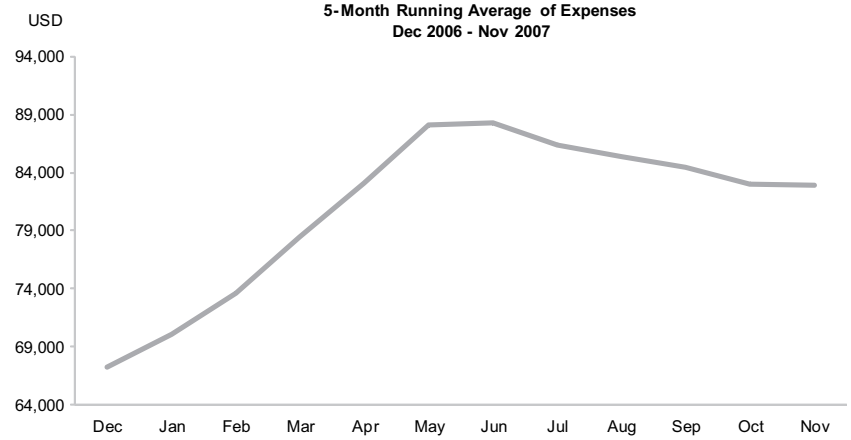


Figure 7.36

The graph on the previous page displays each month's value as the average (mean) of that particular month and the four preceding it. It is often appropriate to examine time series from a smoothed (high-level) and an actual (low-level) value perspective at the same time, as shown in the example below. Seeing both perspectives at once can help us avoid reading too much meaning into either one.

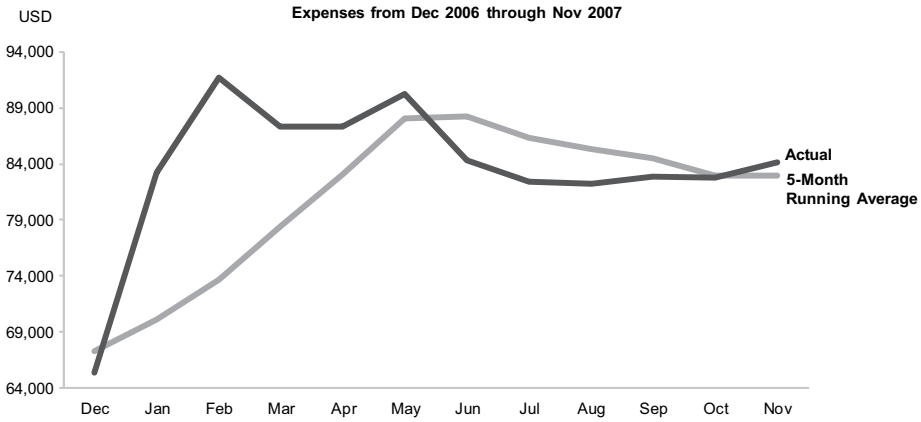


Figure 7.37

Omitting Missing Values from a Display

There's a difference between a value of zero and a value that's missing from the data. The following graph suggests that no employees worked for the company during the month of July.

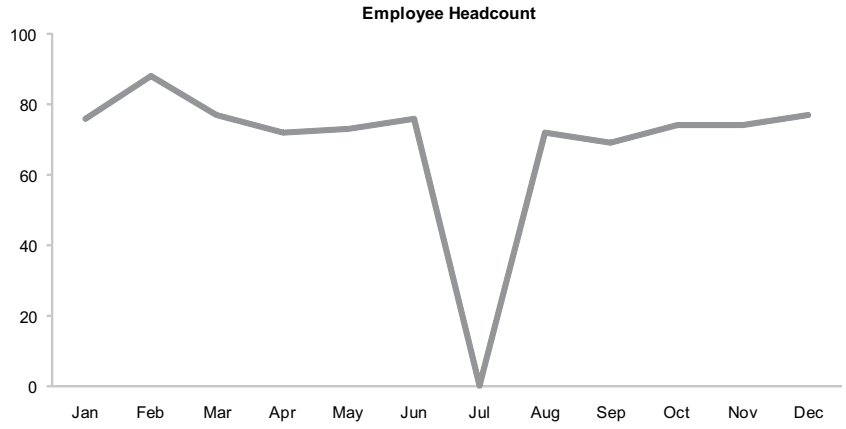


Figure 7.38

It is unlikely that everyone left the company in July and then staffing returned to its previous level in the month of August. Rather, July's employee count is missing from the data, and the graph displayed this omission as a value of zero. Bad graph! This choice produced a picture that doesn't reflect reality. The best way to handle missing values is to omit them from the graph. This makes the fact that values are missing noticeable, and the meaning obvious. Missing values can be visualized in either of the following two ways:

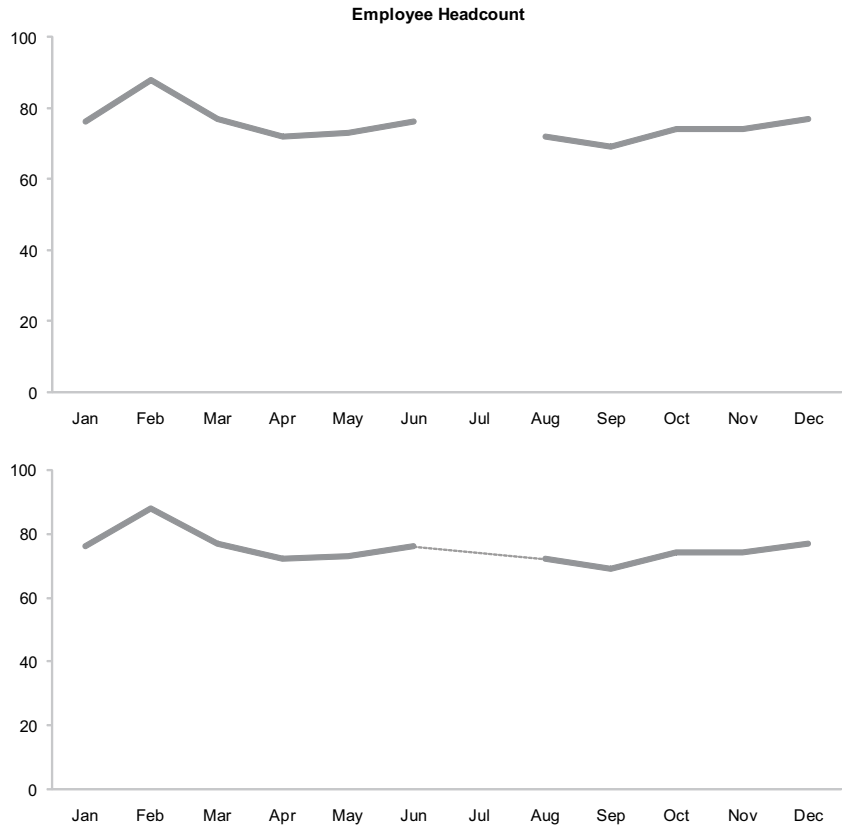


Figure 7.39

When we spot missing values during the course of analysis, we can estimate what's missing or take the time to track down the real values. In the headcount example above, no matter how it's displayed, the value for July is obviously missing because any other interpretation would be absurd, but when we're examining information that at times legitimately includes zeroes, we might not be able to discern the difference between a zero that's real and a value that's missing if both are represented as zero. For this reason, missing values should always be omitted from a graph. If you use software that automatically treats missing values in a graph as zeroes, let the vendor know that this is a bug.

Optimizing a Graph's Aspect Ratio

The aspect ratio of a graph is the ratio of the length of the X-axis to the length of the Y-axis. For example, if the plot area of a graph is exactly as tall as it is wide, it has an aspect ratio of 1 to 1 (sometimes written as 1:1). The aspect ratio of a time-series graph usually works best when it is wider than it is tall. In the example on the following page, the same values are displayed with a 1:1 aspect ratio above and a 2:1 aspect ratio below.

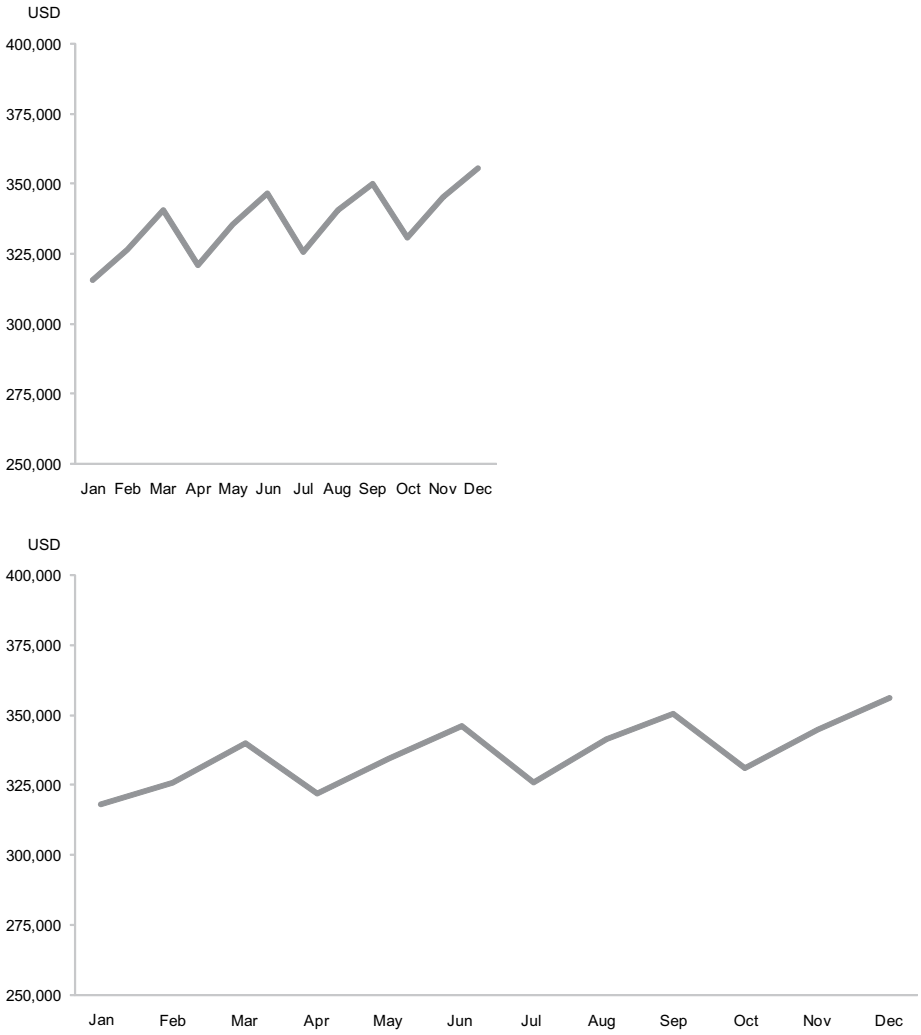


Figure 7.40

In the second graph, we can see that sales rise gradually in each quarter, with a steep decline in the first month of each new quarter. This pattern is much harder to discern in the upper graph, however, because of its aspect ratio.

Despite the usual advantage of making time-series graphs wider than they are tall, no single aspect ratio is always best. The choice of aspect ratio depends on what you're trying to see. It is sometimes worthwhile to experiment with the aspect ratio to see if something meaningful comes to light that wasn't noticeable before. William Cleveland took the time to test various aspect ratios and found that it is often helpful to set them so that the patterns we're focusing on have slopes that are approximately 45°. This is because a 45° slope is easier to see and interpret than one that is flatter or steeper. Cleveland explains the reasons:

> *If the aspect ratio of a display gets too big, we can no longer discriminate two positive slopes or two negative slopes because the orientations get too close. A similar statement holds when the aspect ratio is too small…The orientations of two line segments with positive slopes are most accurately estimated when the average of the orientations is 45°, and the orientations of two line segments with negative slopes are most accurately estimated when the average of the orientations is –45°…The 45° principle applies to*

the estimation of the slopes of two line segments. But we seldom have just two segments to judge on a display, and the aspect ratio that centers one pair of segments with positive slopes on 45° will not in general center some other pair of segments with positive slopes on 45°. Banking to 45° is a compromise method that centers the absolute values of the orientations of the entire collection of line segments on 45° to enhance overall estimation of the rate of change.[2]

2. *The Elements of Graphing Data*, William S. Cleveland, Hobart Press, 1994, pp. 252-254.

As long as we're relying on our eyes to estimate the optimal aspect ratio, it isn't necessary to follow Cleveland's suggestion precisely. Tufte offers a practical solution for time-series displays: "Aspect ratios should be such that time-series graphics tend toward a lumpy profile rather than a spiky profile or a flat profile."[3] The middle graph below illustrates the lumpiness that Tufte advocates, in contrast to the examples of the flatness in the bottom graph and spikiness in the top graph.

3. *Beautiful Evidence*, Edward R. Tufte, Graphics Press, Cheshire CT, 2006, p. 60.

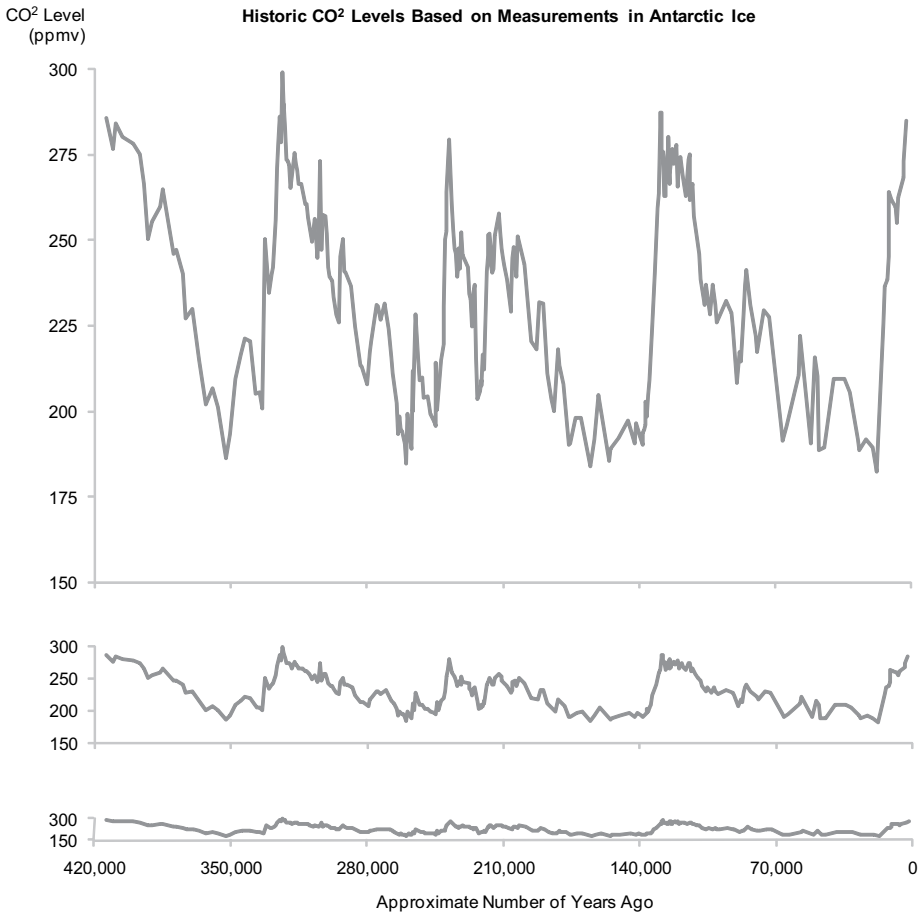


Figure 7.41. Data source: National Climatic Data Center website, based on measurements published by Petit et al., 1999.

Today, to achieve optimal slope, we must manually adjust a graph's aspect ratio, using our eyes alone to roughly determine what's most effective unless we happen to use one of the few products that includes automated banking to 45° algorithms, such as the "R" and "S" languages, which have supported this for many years. Recent work has been done by Jeffrey Heer and Maneesh Agrawala at the University of California, Berkeley to develop even better algorithms for

banking to 45° that could be incorporated into software to improve this process. The option of simply turning on a "banking to 45°" feature in software and having it do the work for us, faster and more accurately than we could possibly do ourselves, is one that I'll welcome with enthusiasm.

Jeffrey Heer and Maneesh Agrawala, "Multi-Scale Banking to 45°," *IEEE Transactions on Visualization and Computer Graphics*, Vol. 12, No. 5, Sept/Oct, 2006.

Using Logarithmic Scales and Percentages to Compare Rates of Change

As we discussed in *Chapter 4: Analytical Interaction and Navigation*, it is natural, when looking at the time-series graph below, to assume that the blue line increased at a faster rate than the brown line.

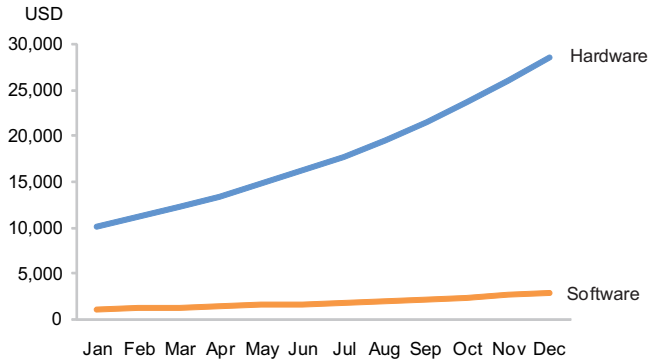


Figure 7.42

In fact, they both increased at precisely the same 10% rate. A 10% increase starting from $1,000 amounts to $100, while a 10% increase starting from $10,000 amounts to $1,000. In a graph with a standard linear scale, the slope of a line that increased by $100 is less steep than one that increased by $1,000. This does not hold true, however, for a graph with a logarithmic (log) scale. The graph below displays the same data, this time using a log scale. Now, equal rates of change appear as equal slopes, no matter how much the actual values are or how great the difference between them.

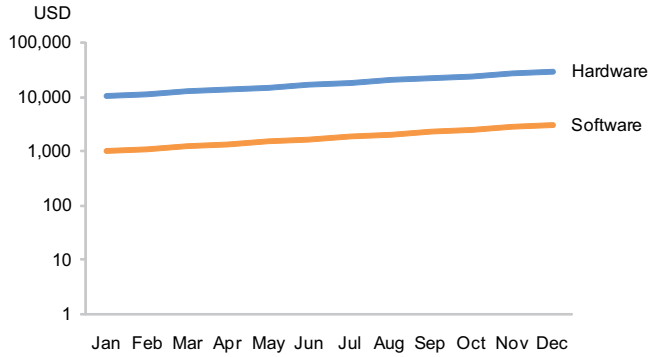


Figure 7.43

The next graph illustrates this from a different perspective. Using a standard linear scale, this graph contains two lines that exhibit precisely the same visual patterns and slopes, which makes it appear that their rates of change were the same.

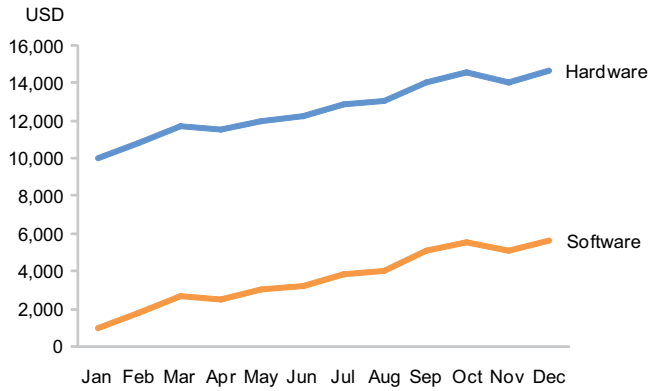


Figure 7.44

The graph below uses a log scale to display the same data, which reveals that the rates of change for hardware and software were, in fact, quite different.

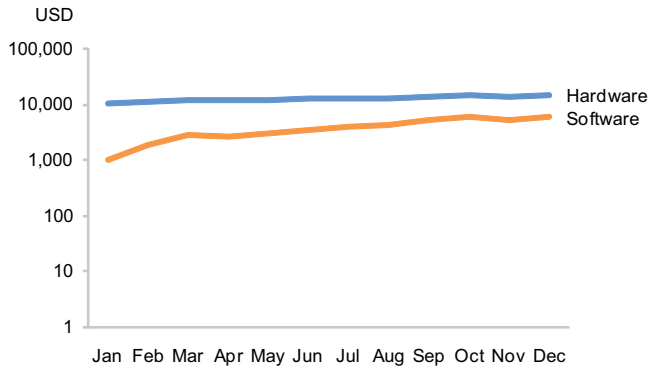


Figure 7.45

Another way to compare rates of change is to graph the rates directly, expressed as the percentage difference between each value and the next. To see how this works, let's begin with a regular graph that compares hardware and software sales throughout a single year.

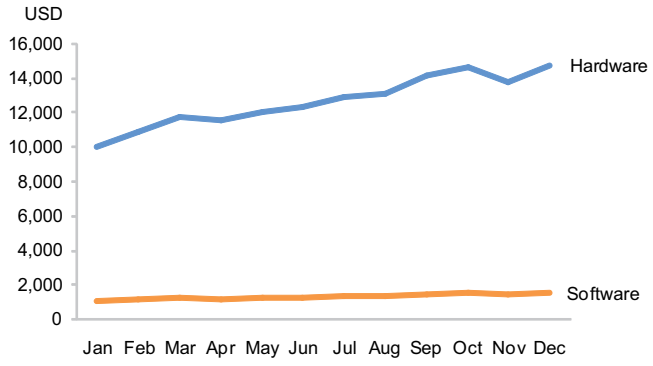


Figure 7.46

As we've learned, it's difficult to compare rates of change using a standard linear scale in this manner. Rather than switching to a log scale, this time let's graph the rates of change directly. The two graphs on the following page display hardware and software sales separately as the percentage change from each month to the next.

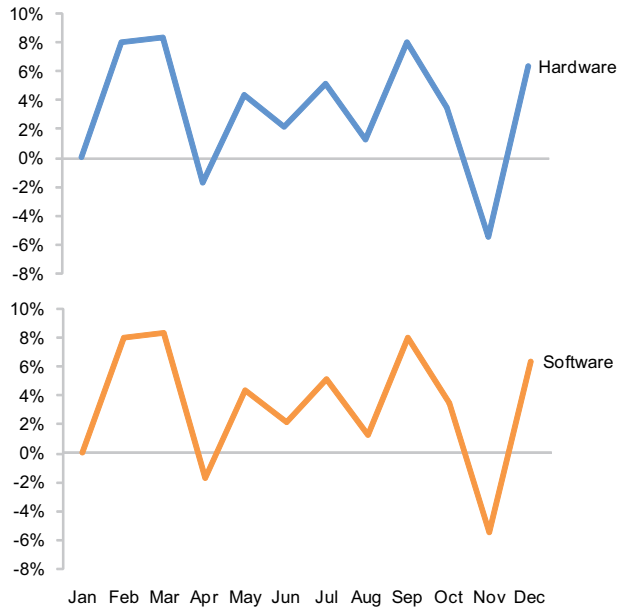


Figure 7.47

As you can see, hardware and software sales exhibited the exact same rates of change from month to month throughout the year. I displayed them in separate graphs only because, had I used a single graph, the two lines would have occupied the same exact space, causing one to be completely hidden by the other. In the next example, rather than graphing the percentage change from one month to the next, I display each month as the percentage difference from a single baseline month, in this case January, the first month of the year. Once again, we can see that the patterns and magnitudes were precisely the same.

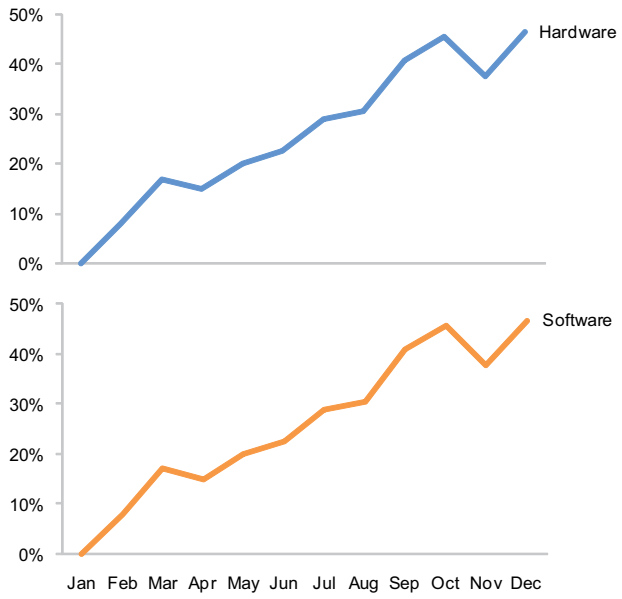


Figure 7.48

Overlapping Time Scales to Compare Cyclical Patterns

We can strengthen our ability to detect and compare cyclical patterns stretching across multiple cycles in a line graph by displaying each cycle as a separate line. As you can see in the following two graphs, the cycles that are difficult to compare in the top graph, are much easier to compare in the bottom graph.

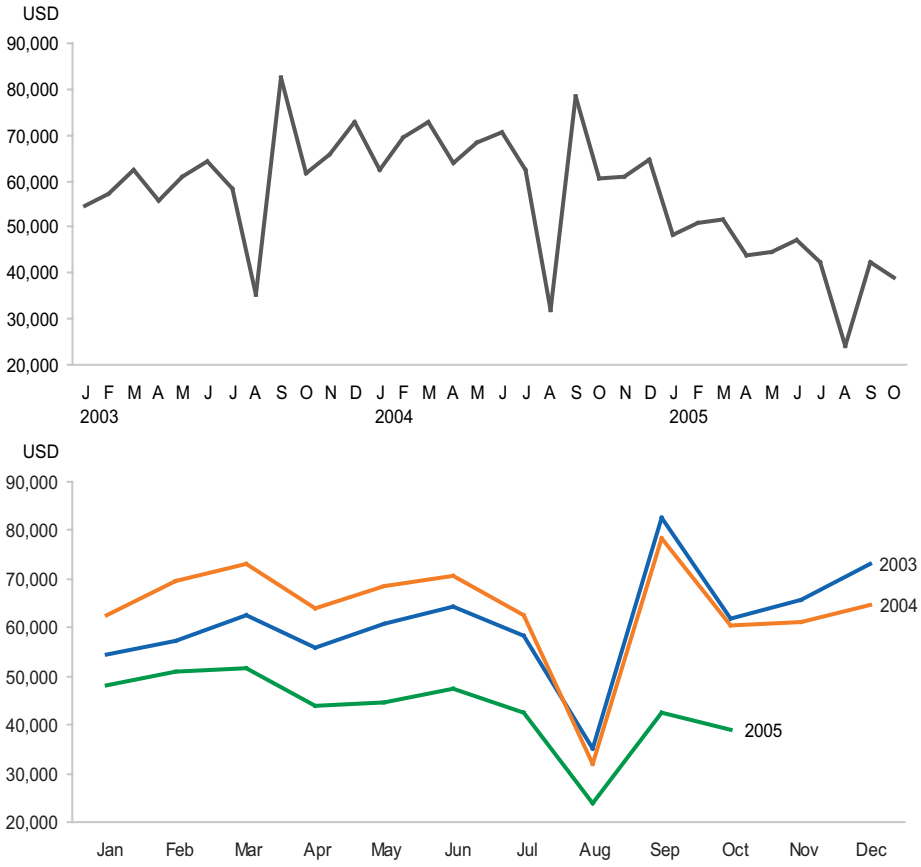


Figure 7.49

This type of graph is particularly easy to create when using Excel, simply by treating each year as a separate data set.

Using Cycle Plots to Examine Trends and Cycles Together

You might have noticed that the technique we just covered in the section above makes it easy to compare cycles to one another but does not allow us to see trends that extend across multiple cycles. A display that makes it possible to both compare cycles and see trends extending across multiple cycles would be useful. The line graph on the next page displays 56 days' worth of sales, which gives us a sense of weekly cycles, but to understand the cycles clearly, we need a different view.

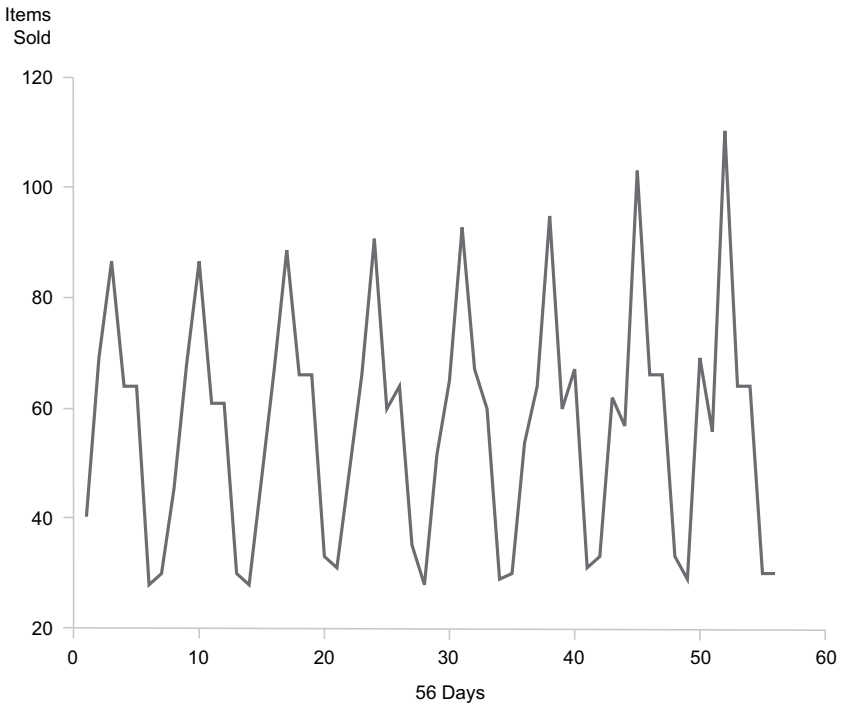


Figure 7.50

The graph below displays the average sales per day of the week for these same eight weeks.

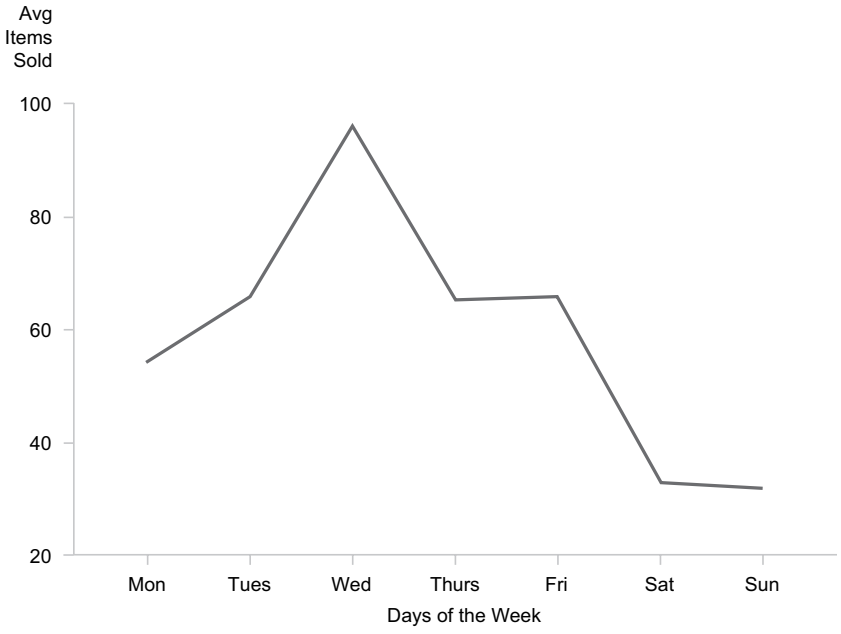


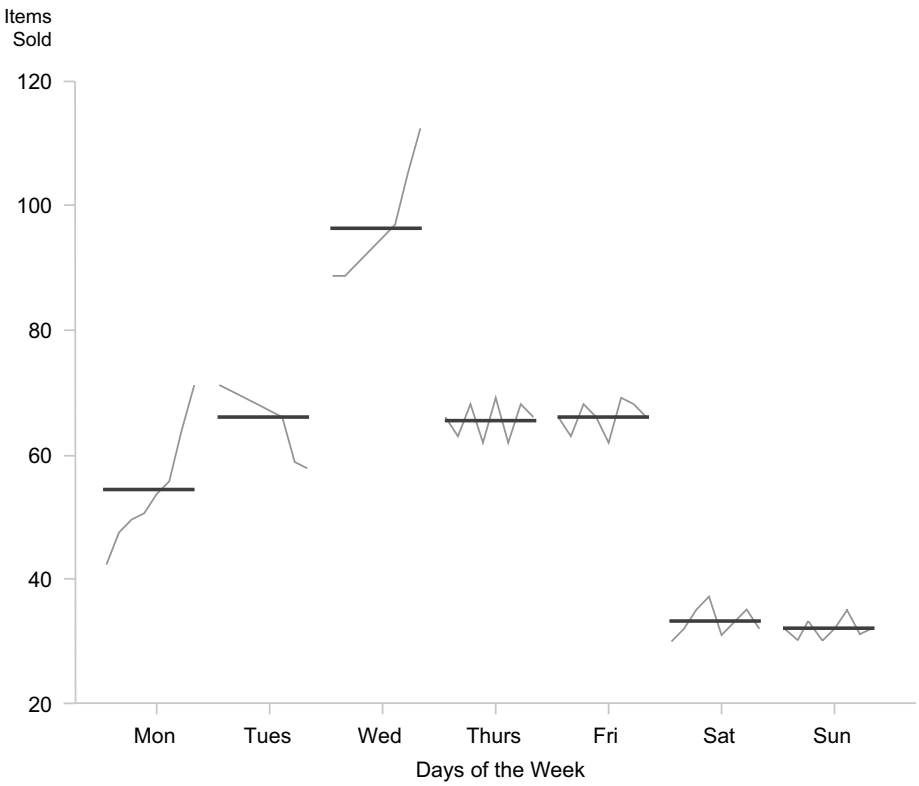
Figure 7.51

We can now see what the overall weekly pattern is for the eight-week period, but we've lost sight of the variation from week to week. In the 1970s, Cleveland, Dunn, and Terpenning developed the *cycle plot*, which can be used to solve this problem.

Cycle plots allow us to see two fundamental characteristics of time-series data in a single graph:

- The overall pattern across the entire cycle
- The trend for each point in the cycle across the entire range of time

Here are the same weekly values that were displayed in the previous graph, this time displayed in a cycle plot:



William Cleveland, Douglas Dunn, and Irma Terpenning, "The SABL Seasonal Analysis Package—Statistical and Graphical Procedures," Bell Laboratories, Murray Hill NJ: Computing Information Service, 1978. This paper was brought to my attention by Naomi B. Robbins in the article, "Introduction to Cycle Plots," *Visual Business Intelligence Newsletter*, Perceptual Edge, Berkeley CA, 2008. Most of the examples of cycle plots shown here were derived from examples that Robbins created for the article.

Figure 7.52

In this cycle plot, the typical weekly pattern is formed across the entire graph by the means (averages) for each day of the week, which are encoded as short, straight horizontal lines. The actual values for any given day of the week across the entire range of time are displayed by each of the small curvy lines. These begin with the value for that particular day of the week during the first week and continue with a value for each week until the last. By looking at the weekly values for Tuesday and comparing them to Monday, we can now see that values consistently increased on Mondays during this eight-week period and consistently decreased on Tuesdays. The two values that stand out as the lowest on Tuesdays occurred during the last two weeks of the period. We can also see that Wednesday also consistently increased, but the other days of the week went up and down without exhibiting a predominant trend.

The ability to summarize cycles and view longer trends without shifting from graph to graph can lead us to insights that we might not otherwise discover. It's useful to have the option of connecting the data points across the graph for any

single cycle (such as any one day of the week in the example above) or across the mean values for all the cycles. In the example below, I've connected the mean values, which makes it easier to see the weekly trend based on the means of each day of the week for the entire 56-day period. Unfortunately, this option doesn't seem to be available in any software that I've seen.

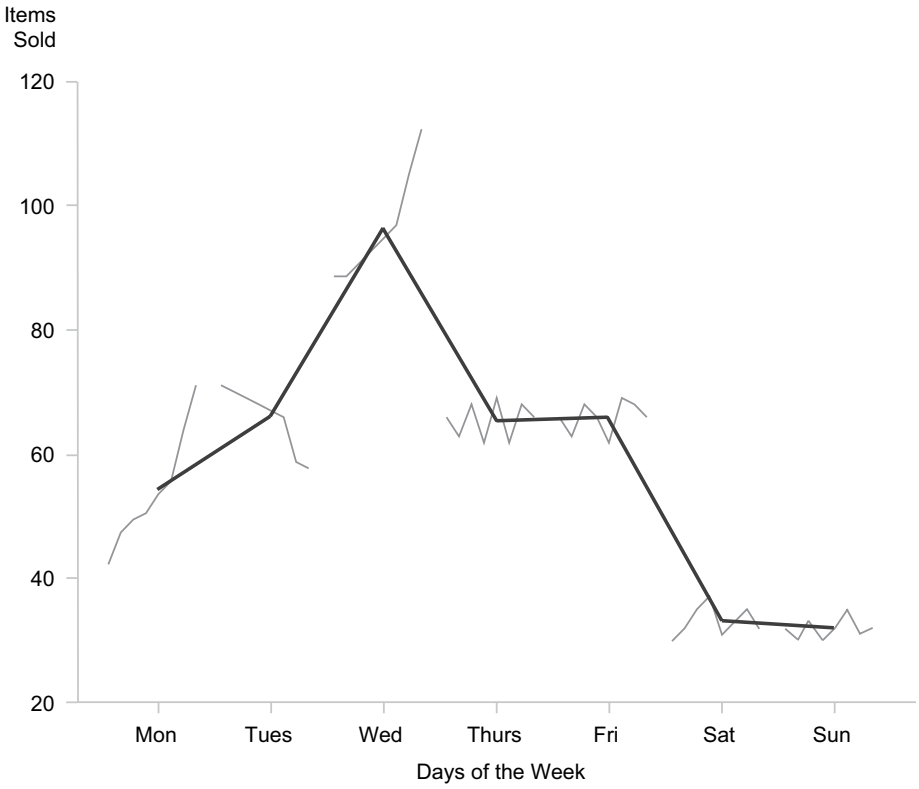


Figure 7.53

Some software products support cycle plots as a special form of line graph. The following example was produced using Tableau Software:

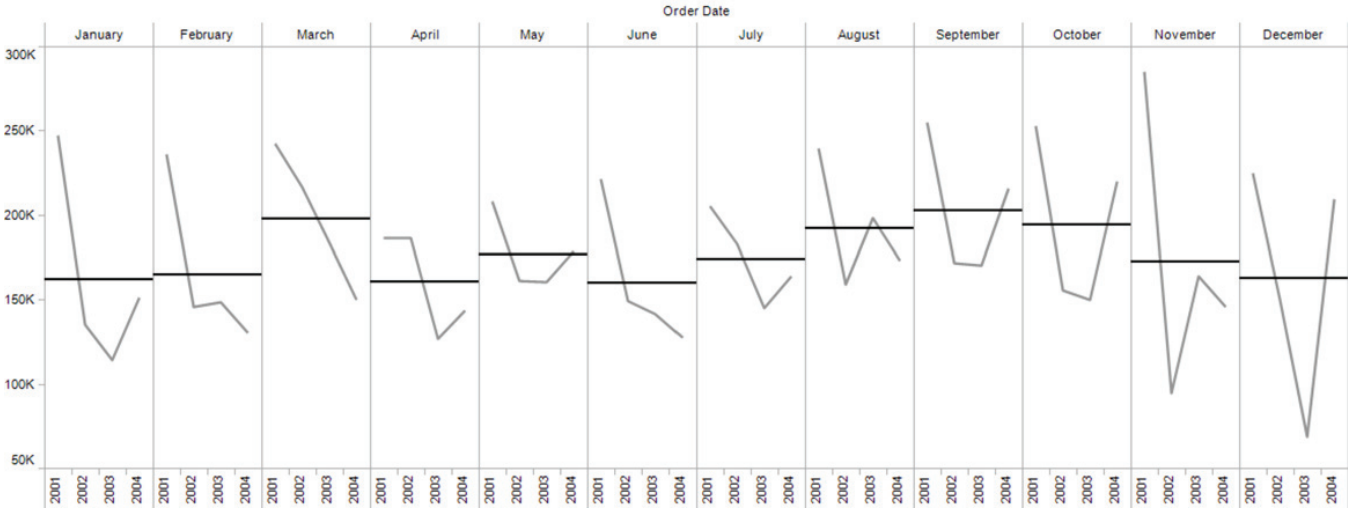


Figure 7.54. Created using Tableau Software

All I had to do to shift from a normal line graph to a cycle plot was to reverse the order of the "month" and "year" fields when I constructed the graph, placing month before year, which caused the years (2001-2004) to be grouped within the months. In other words, it took no longer to construct this cycle plot than it did to construct the normal line graph, and I could quickly and easily switch back and forth between the two simply by reversing the order of the years and months.

Even if you don't have a product like Tableau Software, you can produce cycle plots in Excel with some time and effort. The example below was produced by first creating a single graph for January values (one graph per year from 1993 through 2005), then copying and pasting that graph 11 times to create the others, and finally by selecting the appropriate source data for each month.

The information for this example was acquired from Kelly O'Day of www.ProcessTrends.com. O'Day has developed a procedure that can be followed using Excel to produce a cycle plot of these real estate listings as a single graph, which can be downloaded from his website. It requires a bit of work to format the data and follow the process.

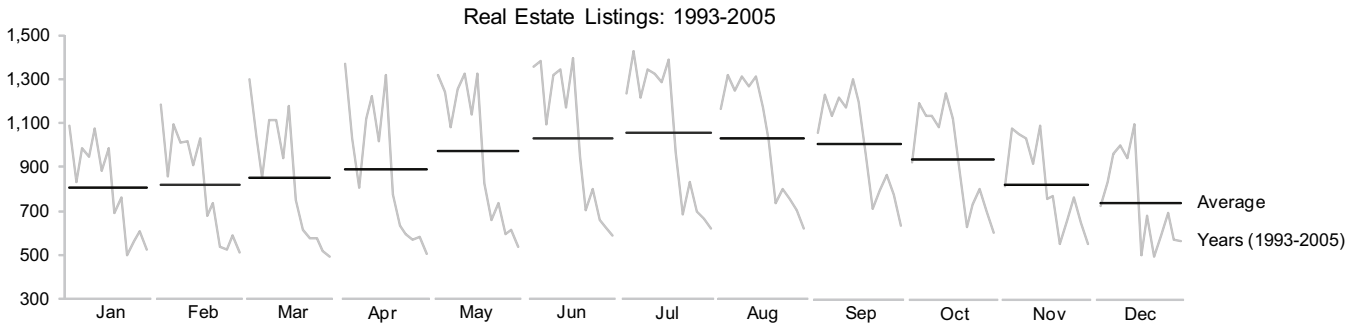


Figure 7.55

Combining Individual and Cumulative Values to Compare Actuals to a Target

Often when we monitor a time series, we assess how well things are going by comparing actual values to targets. If we're tracking expenses on a monthly basis, but the target that's used to judge performance is for the year as a whole, a graph that displays each month's expenses during the course of the year, such as the graph below, would make it hard to compare year-to-date performance to the target.

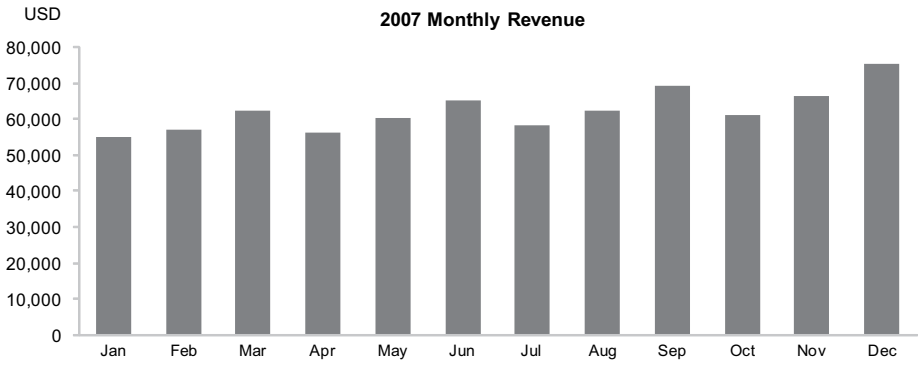


Figure 7.56

We could create a separate graph with just two values—the actual year-to-date expenses and the annual target—but what if we don't want to lose sight of monthly expenses? A simple solution involves a combination bar and line graph,

with a bar for each month's expenses, a line to display cumulative year-to-date expenses per month, and a reference line to mark the target, as illustrated below.

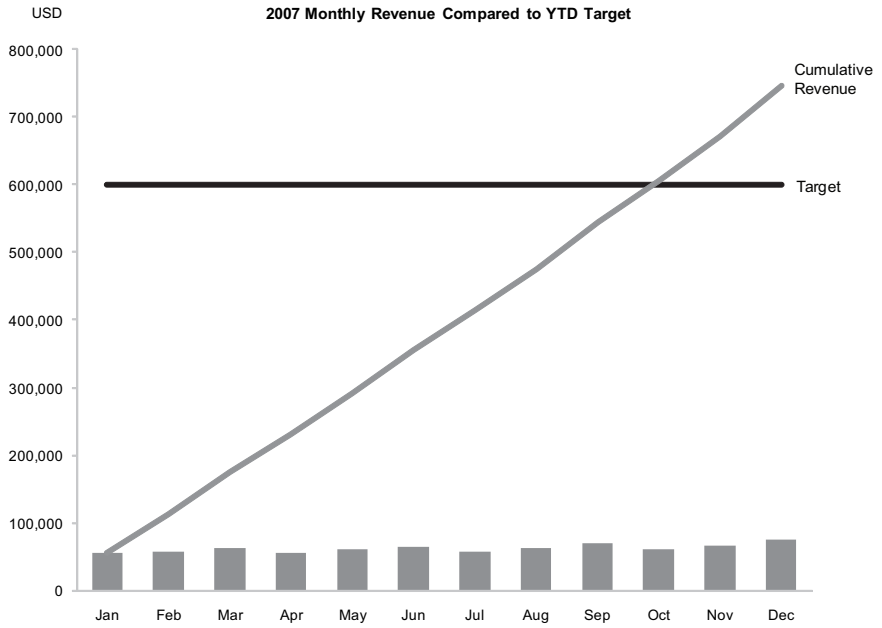


Figure 7.57

A graph of this type is difficult to use to precisely compare individual bars because, when we scale the graph to accommodate the cumulative values, the bars become too short in comparison. If we want to keep the bars longer to eliminate this problem, we can display the data in two graphs, arranged one above the other: a bar graph for monthly values and a line graph for year-to-date cumulative values compared to the target. As illustrated below, we can scale each graph independently, thereby getting the best of both worlds.

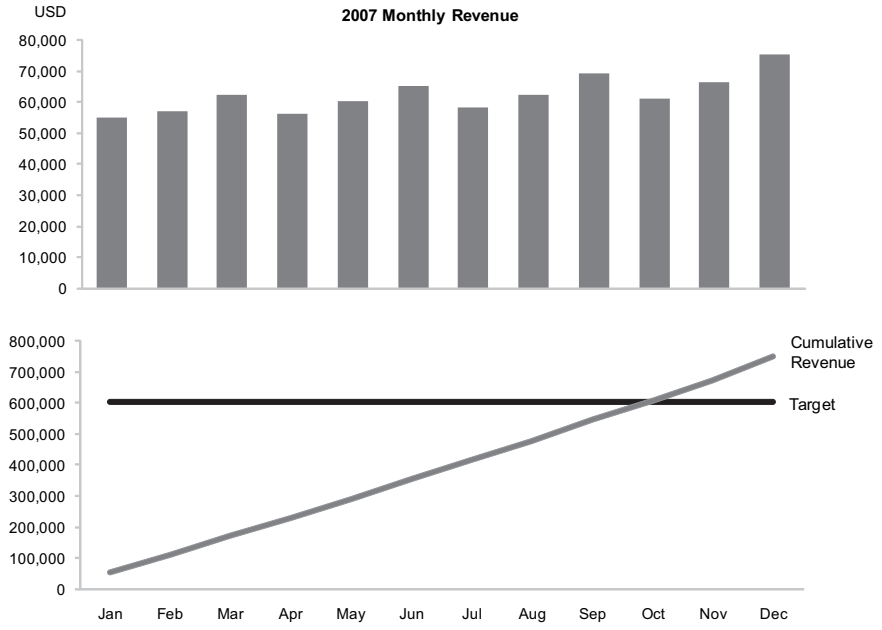


Figure 7.58

Shifting Time to Compare Leading and Lagging Indicators

It is sometimes useful to examine how one variable (the independent variable) affects another (the dependent variable). When we do this in the context of time, it's sometimes difficult to see how the dependent variable is affected by the independent variable if there is a lag in time between the cause and the effect. The following example illustrates this difficulty.

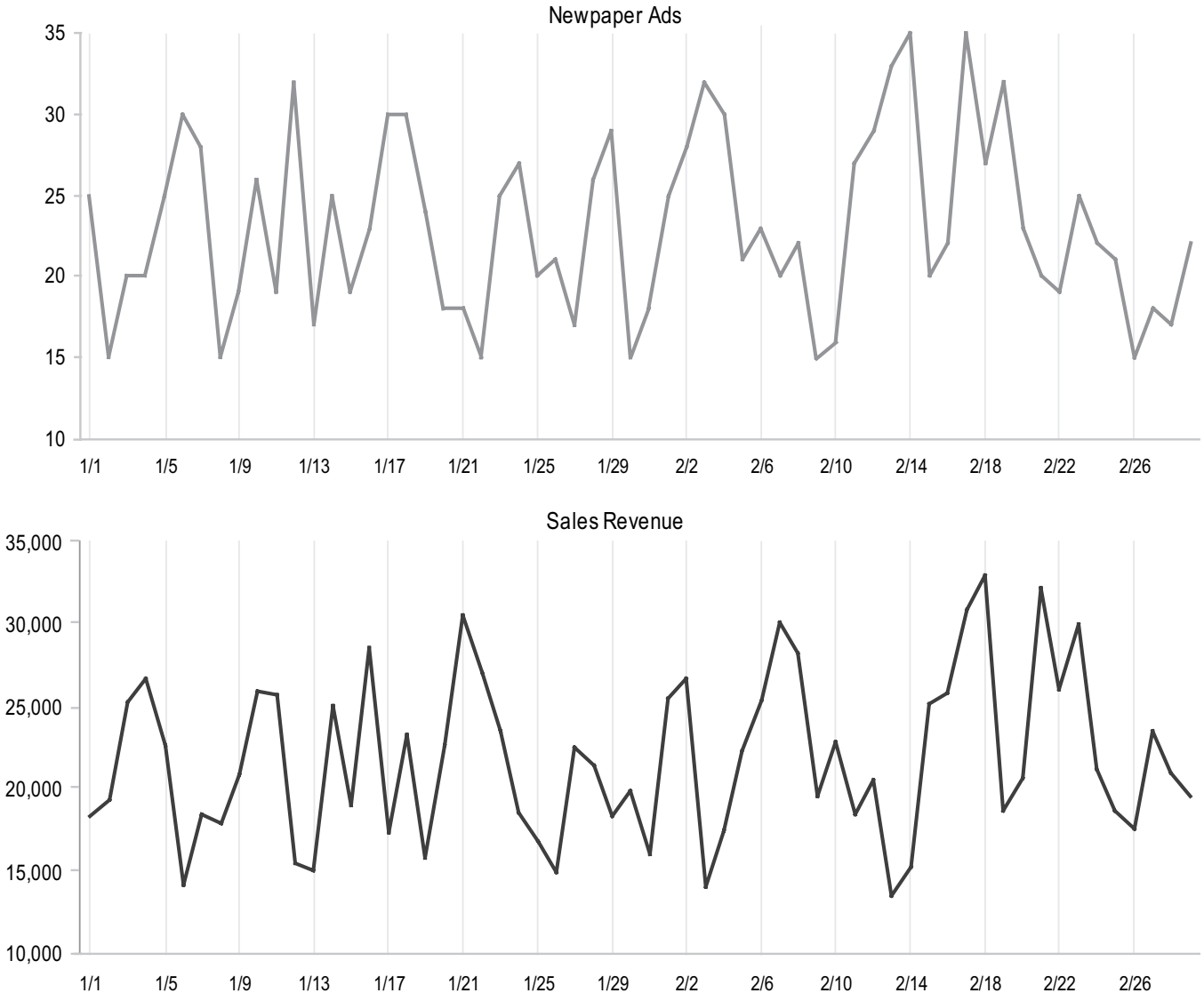


Figure 7.59

Using this example, imagine that we know that a relation exists between newspaper ads and resulting sales such that a greater number of ads results in increased sales four days later. In other words, newspaper ads are a leading indicator of sales revenues, and there is a lag of four days between them. Because of the lag, the up and down patterns formed by the number of newspaper ads don't line up with the related patterns formed by sales revenues. To examine their relationship more closely, we need to align the leading and lagging events.

This example was created using Excel. We can now reposition the sales revenue graph to the left by four days to align the related values, and also move it up a bit to get the two lines closer to one another.

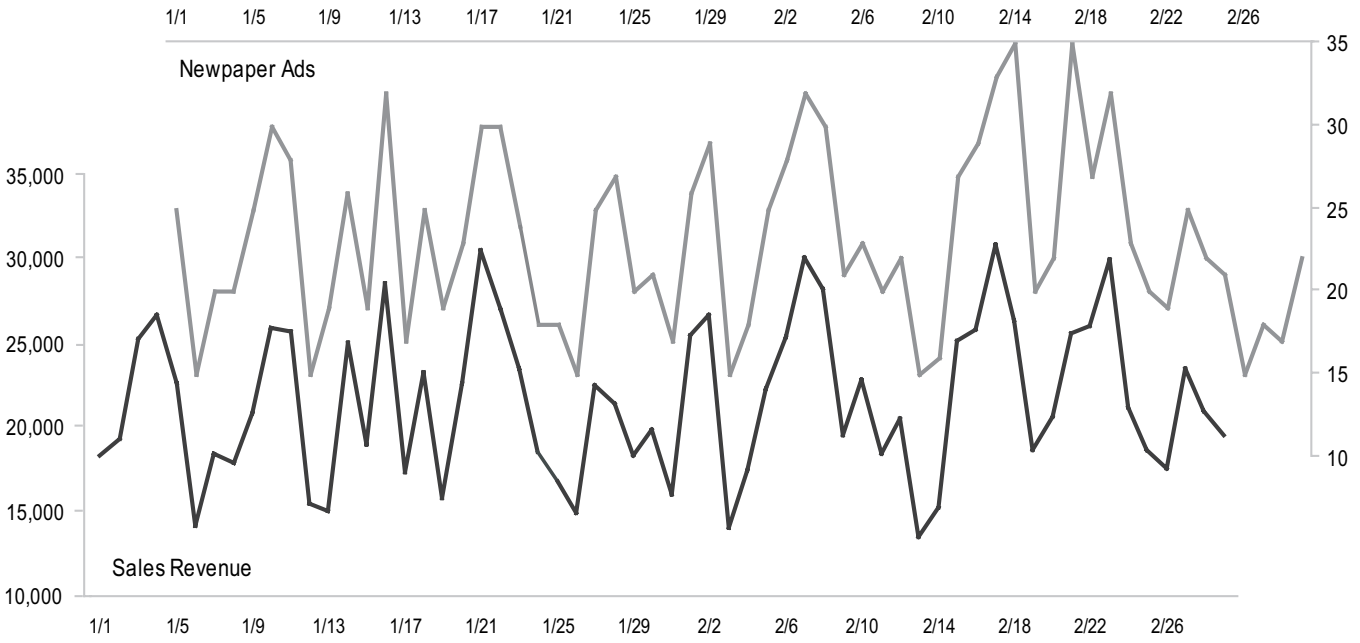


Figure 7.60

Now the leading and lagging values are aligned and can be compared with ease. Doing this in Excel doesn't require anything fancy. Because graphs can be positioned wherever you want them, Excel does a particularly good job of supporting this technique. To place graphs on top of one another so that you can see through the one on top to the one beneath, you simply need to set the chart's background fill color to "none" rather than "white."

A simple solution that I haven't seen so far in a product would involve a feature that allowed us to shift time in a graph to the left or right by any specified amount without affecting time in other graphs that are also on display. Perhaps this feature exists, and I just haven't seen it. If so, the innovative vendor deserves our thanks.

Stacking Line Graphs to Compare Multiple Variables

Often, time-series data sets that are useful to compare can't all reside in a single graph either because they are expressed using different units of measure, or there are huge differences in where they fall along the quantitative scale. For example, we can't compare a product's sales revenues in U.S. dollars to the number of units sold in the same graph without using two scales, which can lead to confusion. Also, it is difficult to compare a product's profits to its average

selling price in a single graph when monthly profits range in the millions of dollars and the average price per product is $500. Scaling the graph to accommodate values in the millions of dollars would cause the average selling prices to barely register as a straight line hugging the bottom of the plot area with no discernable pattern. But it is useful to compare these things, so how can we do it?

This problem can be solved by using a series of graphs arranged above and below one another with the same points in time aligned. Here's an example that compares the number of units sold, revenues, profits, average selling price, and customer satisfaction.

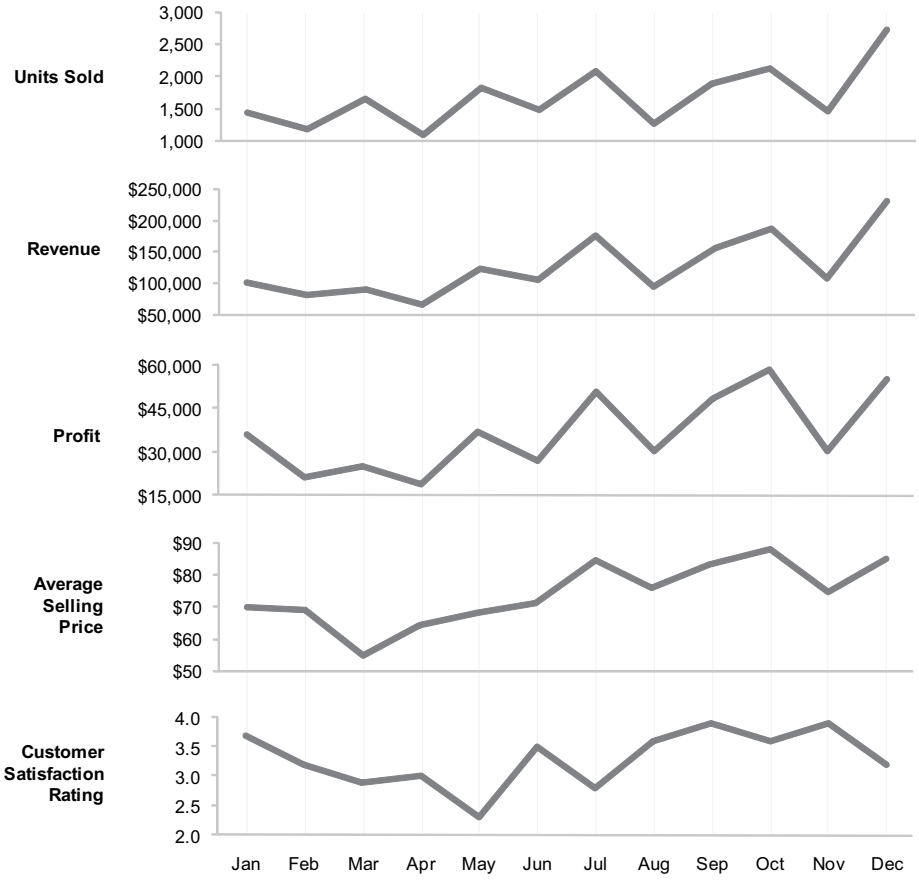


Figure 7.61

When the quantitative scales in graphs are not the same, you can't compare the magnitudes of values in one graph to those in another, but you can compare patterns of change. This technique can be executed in Excel simply by creating separate graphs with the same time scale along their X-axes, and arranging them so that the same points in time are aligned. Other, more powerful products are available for doing this that reduce labor by arranging the graphs automatically.

The following example illustrates how this is done using Tableau Software.

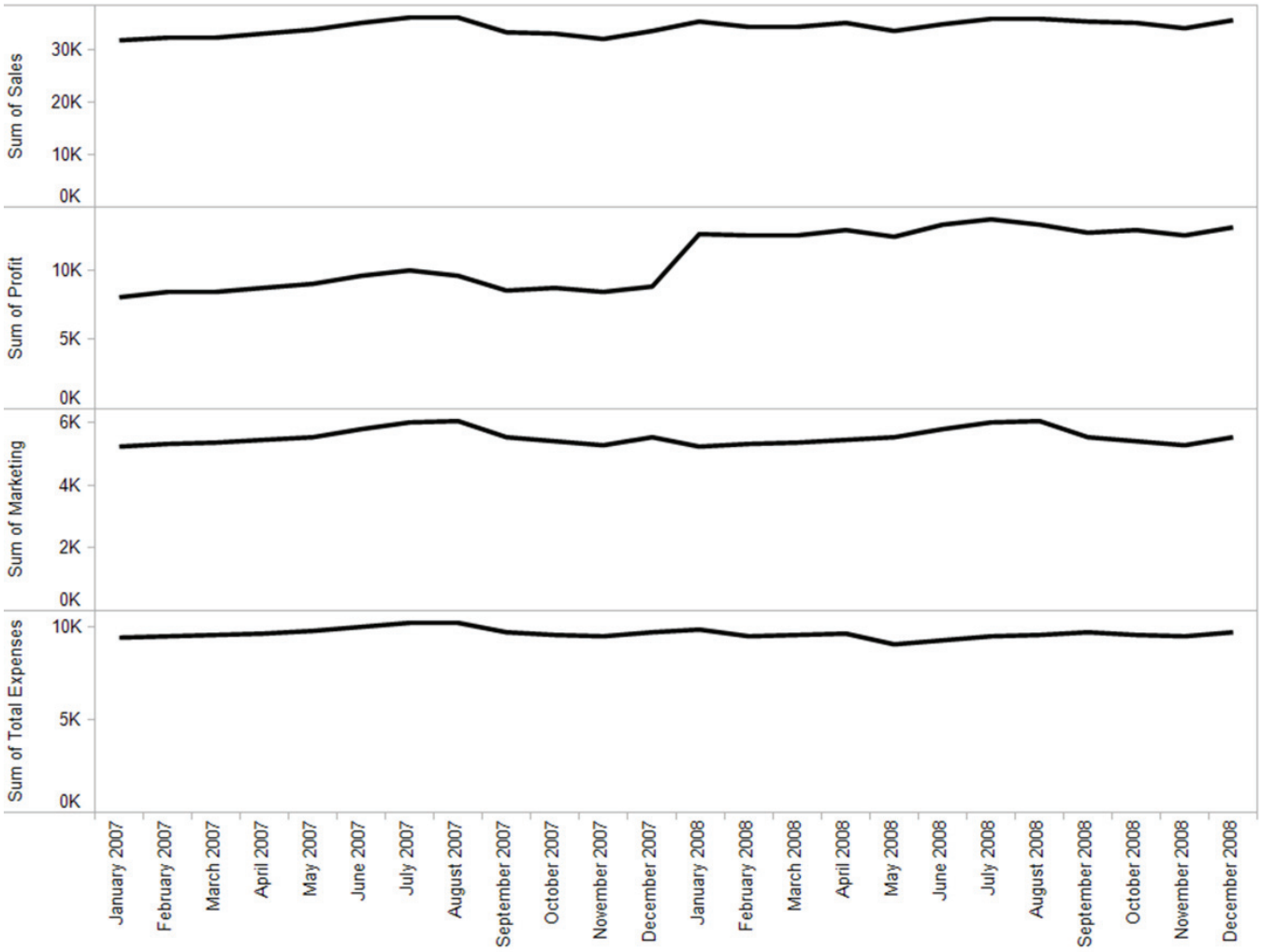


Figure 7.62. Created using Tableau Software

Expressing Time as 0-100% to Compare Asynchronous Processes

Imagine the following situation. You work in the Information Technology (IT) department of a company, and you want to compare costs in person hours associated with 50 projects that IT managed during the past five years. You're interested in finding out whether project costs exhibit particular time-based patterns, such as high costs during the start-up phase, increasing rates of costs near the end, or other patterns that you have yet to imagine. The problem that prevents you from analyzing these costs as you would any other time series is the fact that the 50 projects did not all start at the same time, nor did they all last the same length of time. In other words, they were asynchronous. What can you do?

One answer involves making starting time, ending time, and duration consistent for all projects. This can be done by expressing each project's duration as a percentage, beginning at 0% and ending at 100%, no matter when the project began, when it ended, or how long it lasted. This approach makes it possible to compare what's happening at the beginning of each process, at the end of each process, halfway through each process, 90% through each process, and so on, despite their asynchronous nature. I ran across this solution in a research project done by the Human-Computer Interaction Lab (HCIL) at the University of Maryland, which produced a software application called *TimeSearcher*.

The example below, prepared using TimeSearcher 2, compares the bid prices (top graph) and velocities (bottom graph) of 227 eBay auctions (one line per auction). Velocity is the rate at which bids were being made.

TimeSearcher 1 and *TimeSearcher 2* were developed at the University of Maryland (www.cs.umd.edu/hcil/timesearcher). TimeSearcher 1 was developed under the direction of Ben Shneiderman by Harry Hochheiser. TimeSearcher 2 is described in the following research paper: Aleks Aris, Ben Shneiderman, Catherine Plaisant, Galit Shmueli, and Wolfgang Jank, "Representing Unevenly-Spaced Time Series Data for Visualization and Interactive Exploration." *Proceedings of the International Conference on Human-Computer Interaction*, 2005, pp. 835-846.

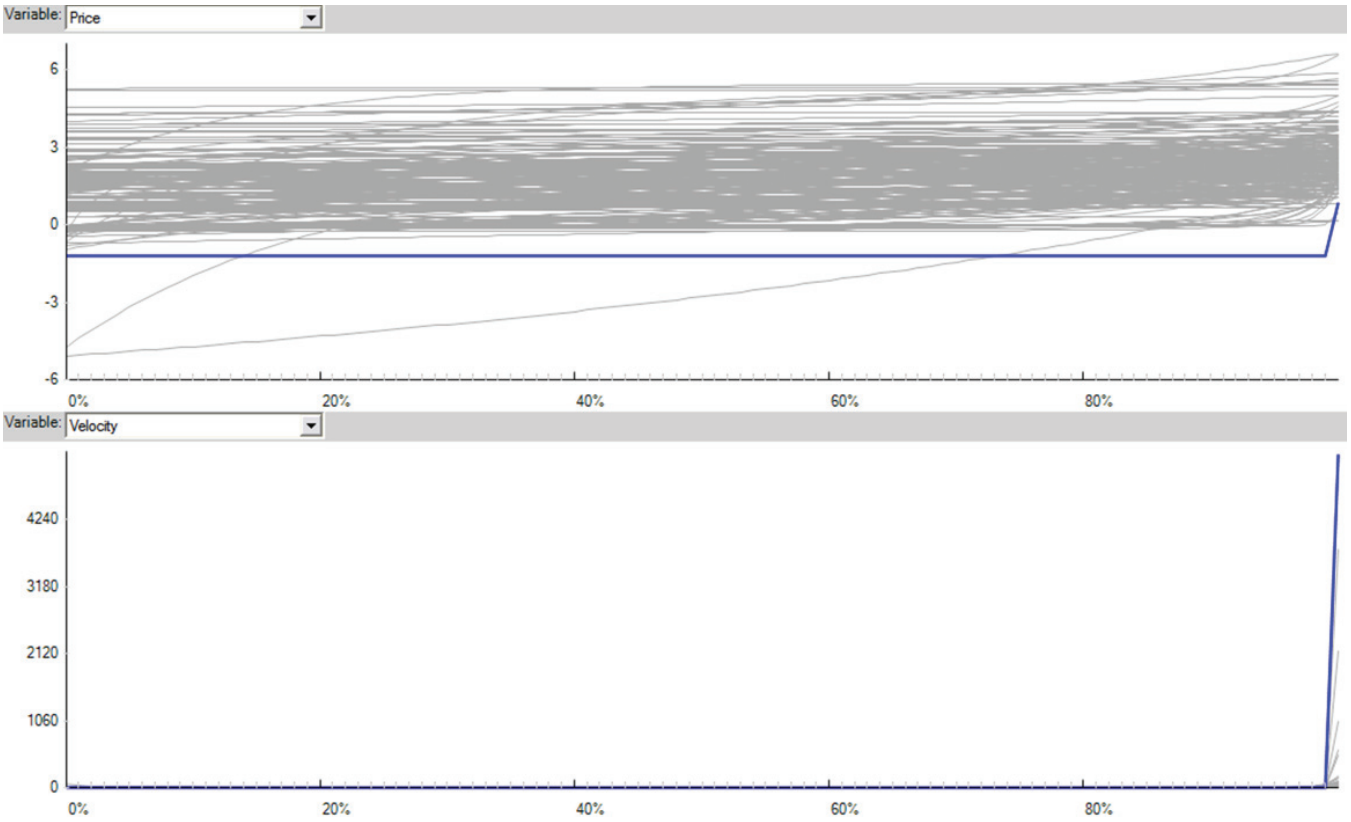


Figure 7.63. Created using TimeSearcher 2

These auctions started on different days and lasted different numbers of days, yet their time-based patterns can be meaningfully compared in this manner. The lines that are highlighted in blue represent an auction that I found interesting because it exhibited the greatest surge of bid velocity of all, revealed by steep line at the right end of the bottom graph. As you can see, the bid price for this auction remained constant until the very end, when the surge in bidding activity produced a slight increase in the price.

In the following example, I highlighted the 10 auctions with the highest final bid prices using the aqua colored rectangle, which TimerSearcher 2 calls a *timebox*, to see whether auctions with high final prices exhibit a particular velocity pattern.

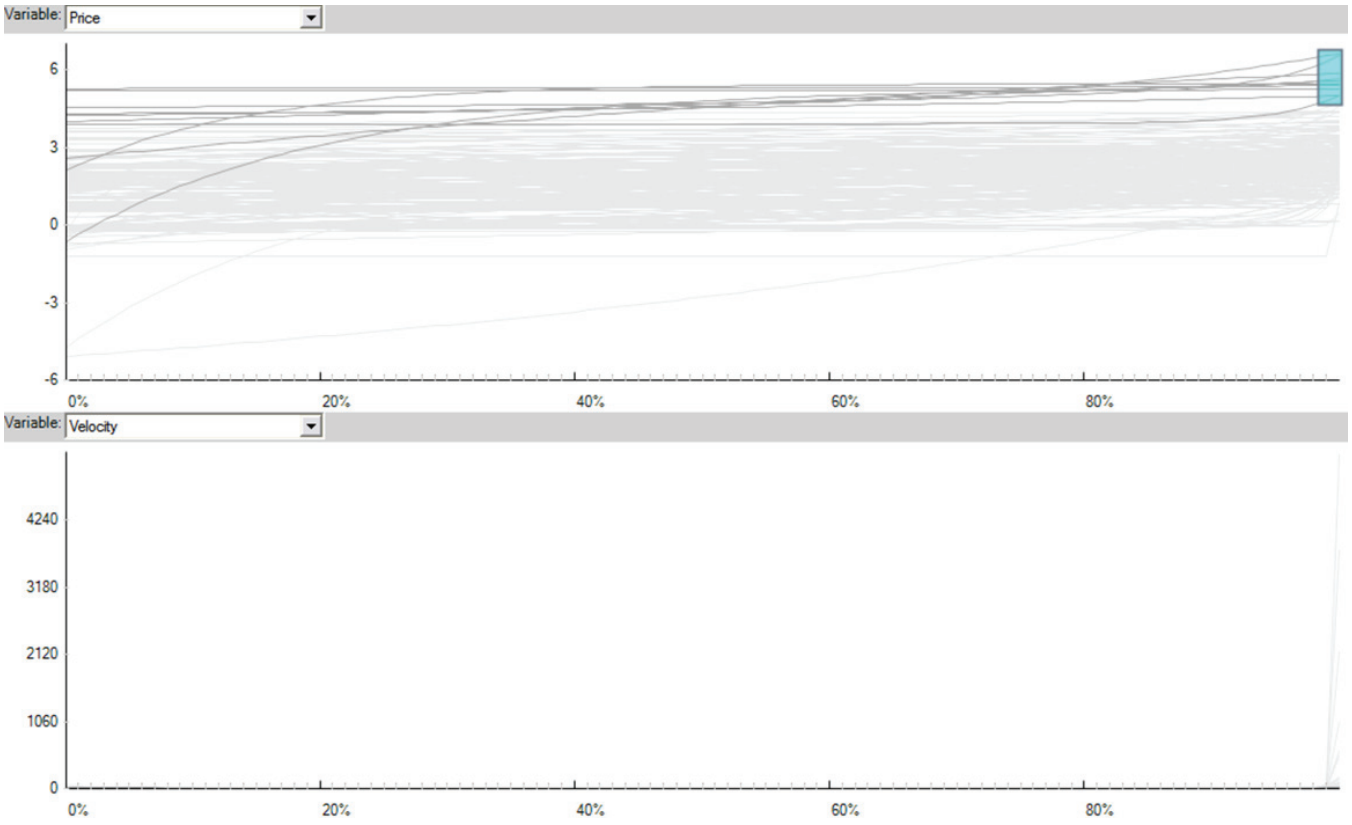


Figure 7.64. Created using TimeSearcher 2

Of the 10 auctions that are highlighted in the price graph (the darker graph lines in the top graph), none is highlighted in the velocity graph, which tells us that none exhibited a significant velocity increase near the end of the auction.

So far, I haven't seen any software that automatically converts time to percentages in the manner described above. TimeSearcher requires that this conversion be done before the program accesses the data. For now, we can do this conversion ourselves. For example, with Excel, we can convert dates to a 100% scale by following a relatively simple procedure, described in *Appendix A: Expressing Time as a Percentage in Excel*. Once this is done for the dates associated with each process that we want to compare, we can use an Excel scatterplot—the version that connects values sequentially with a straight line—to display the data, which I did to produce the following example.

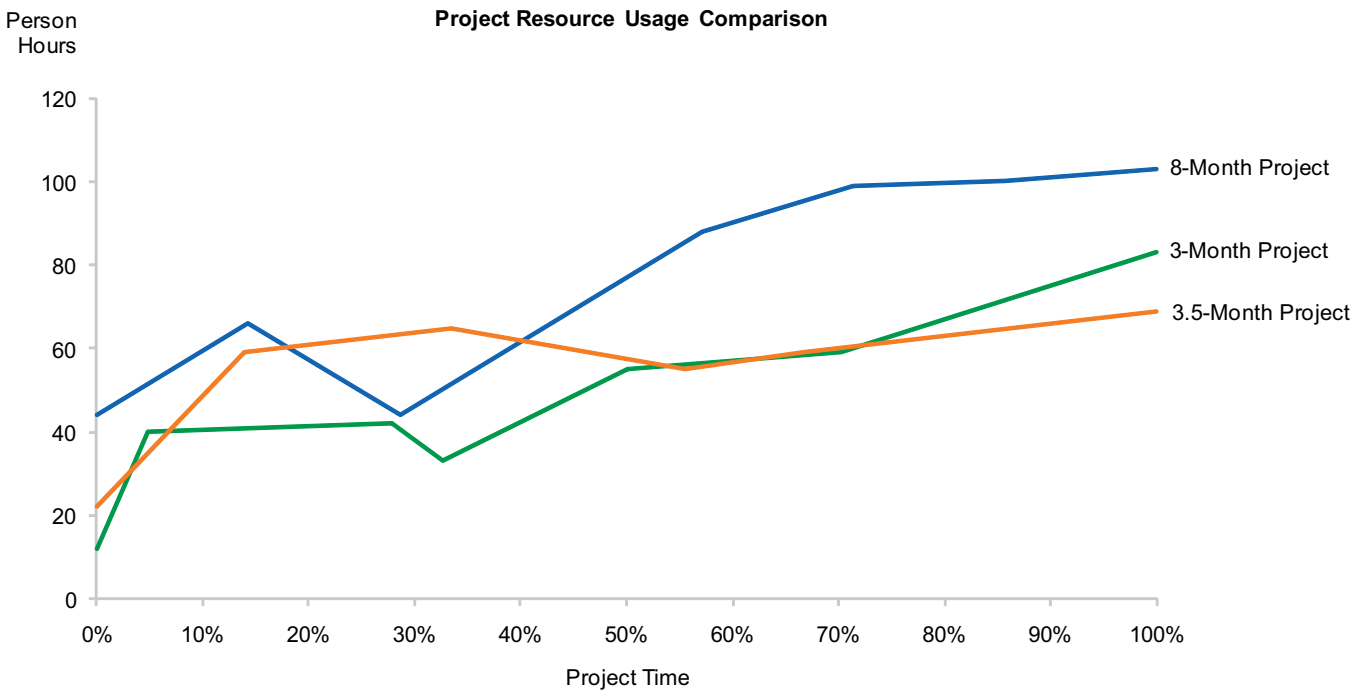


Figure 7.65

Searching for Specified Patterns of Change

The technique of searching for specific patterns was brought to my attention initially by the TimeSearcher applications mentioned above. Wouldn't it be great if you could draw a time-series pattern in the form of a line, such as  and tell your software to find and display every line that exhibits a similar pattern within a specified range of variation? TimeSearcher doesn't do this exactly, but it does let you pick a section of a line in a graph and then instruct the software to find all other patterns throughout the data set that are similar in shape. The ability to draw a pattern and instruct the software to look for it was implemented by Martin Wattenberg in the research project that produced *QuerySketch*, but it has never, to my knowledge, been implemented in commercial software. Unlike most of the analysis techniques that I describe in this book, this one won't be practical until a software product supports it. I hope that by the time you read this, products will save us time by searching massive amounts of data to find specific patterns much faster than we could ever do with our eyes alone.

"Sketching a Graph to Query a Time-Series Database", Martin Wattenberg, Dow Jones / SmartMoney.com, New York NY, 2001; Wattenberg currently works for IBM Research and is responsible for some of the best information visualization research and development being done today.

Maintaining Consistency through Time

I've found that business analysts often ignore two important practices that are necessary to maintain consistency in time-series values, especially when those values extend across several years:

- Adjusting for inflation when examining currency
- Taking into account differences in how the information was collected or defined over time

When we wish to examine and compare monetary values across multiple years, inaccuracies will arise if we fail to account for the changing value of money that results from inflation. This failure can cause us to conclude that a product's sales performance is greater today than it was five years ago, even when its performance has in fact decreased. I believe that adjusting for inflation is seldom done primarily because people simply don't think about it or they assume that it's much harder than it actually is. Indexes that can be used to adjust for inflation are readily available for download from the Internet and can be plugged right into software such as Excel. Instructions for finding these indexes and using them in Excel can be found in *Appendix B: Adjusting for Inflation in Excel*.

In *Chapter 2: Prerequisites for Enlightening Analysis*, I mentioned how important it is to know the pedigree of your data. One reason is that sometimes a particular measure, even of something as common as sales bookings, could have been defined or calculated in the past differently than it is now. Differences like this are especially common if your organization switched between computer systems at some point in the past, and the old system calculated or defined some things differently than the current system does. Another common cause in business is the acquisition of another company whose systems defined measures differently than yours, and the merging of that company's data with yours. For example, if one system defined revenues such that sales taxes were included and the other treated sales taxes as separate from revenues, and information from them was merged five years ago without taking this difference into account, comparing six-year-old revenues to current revenues would be like comparing apples and oranges (or at least McIntosh and Granny Smith apples) unless you adjusted for this difference. If someone else, such as a data warehousing team, takes care of adjustments like these, you're lucky. If not, you ought to make these adjustments yourself.