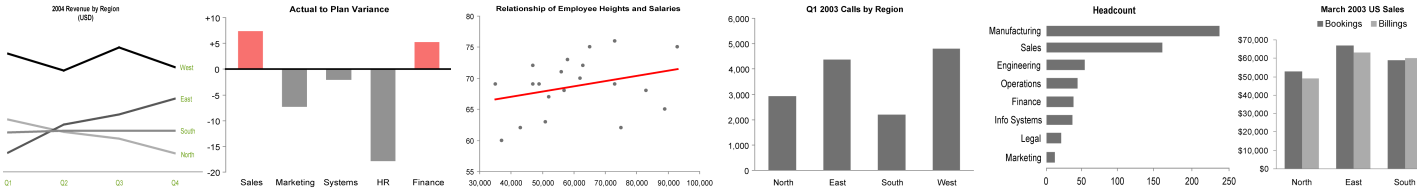


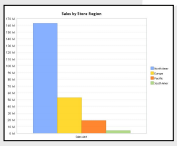
Effectively Communicating Numbers

Selecting the Best Means and Manner of Display

by

Stephen Few
Principal, Perceptual Edge
November 2005

SPECIAL ADDENDUM



Effectively Communicating Numbers
with ProClarity

ProClarity™

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The ability to display data graphically is not intuitive; it requires a set of visual design skills that must be learned. Based on the recent book, *Show Me the Numbers: Designing Tables and Graphs to Enlighten*, this white paper will introduce the best practices in graph design.

No information is more important to a business than quantitative information – the numbers that measure performance, identify opportunities, and forecast the future. Quantitative information is often presented in the form of graphs. Unfortunately, most graphs used in business today are poorly designed – often to the point of misinformation. Why? Because almost no one who produces them, including specialists such as financial analysts and other report developers, have been trained in effective graph design.

This white paper is designed to provide a practical introduction to graph design developed specifically for the needs of business. Following these clear precepts, communicated through examples of what works and what doesn't, you will learn a step-by-step process to present your data clearly and drive your message home.

You Will Learn To:

- Match your message to the right type of display
- Design each component of your graphs so the data speaks clearly and the most important data speaks loudly

ProClarity sponsored this white paper in order to help people understand and design the most effective ways to present quantitative information in general or while using ProClarity business intelligence solutions.

INTRODUCTION

Imagine that it is Thursday afternoon and an email from your boss suddenly appears in your inbox. With a sigh you wonder, "What's Sue want this time?" You open the email and here's what she says:

> *I've interviewed three people for the new Customer Service Manager position and need to summarize their qualifications for Jeff [the big boss]. He wants to choose the best candidate as objectively as possible. After the colossal failure of my last hire, he no longer trusts my instincts. I've attached a spreadsheet that rates each of the candidates according to the six areas of competence that we use for performance reviews (experience, communication, etc.). Please create a report that I can pass on to Jeff that presents me findings. I'll need it on my desk first thing tomorrow.*

Handling requests like this is your job, but you've never before been asked to present the qualifications of potential hires. Not only do you want to impress Sue, but here's a chance to impress the big boss as well. Obviously, you've got to pull something great out of your hat—not any old table or graph will do. You run through the list of possibilities and select what you hope will win the day. Here's what you have waiting on Sue's desk the next morning.

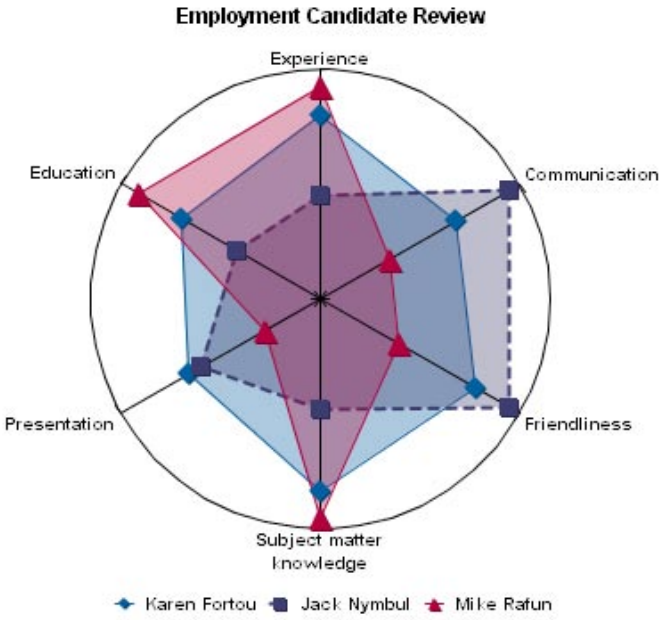


Figure 1: This is an actual example of a software vendor's (not ProClarity) idea of an effective graph.

No run-of-the-mill employee would think to use a radar chart. You figure that a bar chart would have been mundane, but the radar chart shown in figure 1 looks very cool, very cutting edge. At about 8:30 AM you receive another email from Sue. It says "Great Job!!!" following by a smiley face. You begin imagining what you will do with the raise you certainly deserve.

In truth, however, a radar chart is not the best fit for this particular data and purpose. It unnecessarily complicates an otherwise simple message. In this case a plain old table, like the one in table 2, would have communicated much more clearly. It's not fancy, but if the goal is communication leading to understanding, this table works exceptionally well. Jeff, the "big boss," would have no difficultly making sense of it. The three candidates are ranked in the order of their overall qualifications ("Average Rating") from left to right. Comparisons between their qualifications in any single area (for example, "Subject Matter knowledge") can be easily made given this tabular arrangement of the data.

Employment Candidate Review

	Candidates		
Rating Areas	Karen Fortou	Mike Rafun	Jack Nymbul
Experience	4.00	4.50	2.50
Communication	3.50	2.00	5.00
Friendliness	4.00	2.00	4.50
Subject matter knowledge	4.00	5.00	2.50
Presentation	3.00	1.50	2.75
Education	3.50	4.50	2.00
Average Rating	3.67	3.25	3.21

Figure 2

Scenarios such as this are not unusual. Decisions regarding how to display quantitative business data are rarely rooted in a firm understanding of which medium would communicate most effectively. In fact, "effective communication" often fails to even make the list of criteria that are considered. If you don't possess basic "graphicacy" skills—competence in communicating information in graphical form—your decisions about how best to present information are arbitrary and often ineffective—sometimes to the point of misinformation.

Quantitative information—numbers—need never suffer in this way. If you understand them, there are ways to communicate their meaning with exceptional clarity. Back in 1954, when Darrel Huff wrote his book "How to Lie with Statistics," he exposed an insidious problem: the presentation of quantitative information in ways that were intentionally designed to obscure and mislead. This problem still exists today, but a more common problem and one that is much more insidious because it is so seldom recognized, is the unintended miscommunication of quantitative information that happens because people have never learned how to communicate it effectively. Most business graphs that I see fit into this category. They communicate poorly, if at all. You have a chance, however, to become an exception to this costly norm.

Fortunately, the skills necessary to effectively communicate most quantitative business data don't require a Ph.D. in statistics. In fact, they are quite easy to learn, but learn them you must. You must know a little about the ways that quantitative data can be visually encoded in a graph, which type of encoding works best under which circumstances, how to avoid the inclusion of anything visual that distracts from the data, and how to highlight those data that are most important to the message you're trying to communicate. The process of selecting and constructing a graph can be approached as a sequential series of decisions, one at a time. My goal in this white paper is to sequence and describe this series of decisions in a way that not only reveals the right decisions for particular circumstances, but the reasons that they are right so you can apply them with understanding.

This process consists of the following six fundamental stages:

1. Determine your message and identify the data necessary to communicate it.
2. Determine if a table, graph, or combination of both is needed to communicate your message.

The remaining stages apply only if one or more graphs are required.

3. Determine the best means to encode the values.
4. Determine where to display each variable.
5. Determine the best design for the remaining objects.
6. Determine if particular data should be featured above the rest, and if so, how.

GENERAL CONCEPTS AND PRACTICES

Before we dive into the graph design process, there are a few general concepts that you should learn, which apply in all circumstances, beginning with the appropriate uses of tables versus graphs.

TABLES VERSUS GRAPHS

In general, when comparing tables and graphs as means to present quantitative data, neither is better than the other—they are simply different, with different strengths and applications. Let's begin by defining the terms.

Table	Graph
Data are expressed in the form of text (that is, words and numbers, rather than graphically)	Data are expressed graphically (that is, as a picture)
Data are arranged in columns and rows	Data are displayed in relation to one or more axes along which run scales that assign meaning to the values

These differences correspond to different strengths as means to present data.

Tables work best when the display will be used to look up individual values or the quantitative values must be precise. Graphs work best when the message you wish to communicate resides in the shape of the data (that is, in patterns, trends, and exceptions). Take a look at figure 3. This table contains rates, organized by year and month.

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Annual
1990	127.4	128.0	128.7	128.9	129.2	129.9	130.4	131.6	132.7	133.5	133.8	133.8	130.7
1991	134.6	134.8	135.0	135.2	135.6	136.0	136.2	136.6	137.2	137.4	137.8	137.9	136.2
1992	138.1	138.6	139.3	139.5	139.7	140.2	140.5	140.9	141.3	141.8	142.0	141.9	140.3
1993	142.6	143.1	143.6	144.0	144.2	144.4	144.4	144.8	145.1	145.7	145.8	145.8	144.5
1994	146.2	146.7	147.2	147.4	147.5	148.0	148.4	149.0	149.4	149.5	149.7	149.7	148.2
1995	150.3	150.9	151.4	151.9	152.2	152.5	152.5	152.9	153.2	153.7	153.6	153.5	152.4
1996	154.4	154.9	155.7	156.3	156.6	156.7	157.0	157.3	157.8	158.3	158.6	158.6	156.9
1997	159.1	159.6	160.0	160.2	160.1	160.3	160.5	160.8	161.2	161.6	161.5	161.3	160.5
1998	161.6	161.9	162.2	162.5	162.8	163.0	163.2	163.4	163.6	164.0	164.0	163.9	163.0
1999	164.3	164.5	165.0	166.2	166.2	166.2	166.7	167.1	167.9	168.2	168.3	168.3	166.6
2000	168.8	169.8	171.2	171.3	171.5	172.4	172.8	172.8	173.7	174.0	174.1	174.0	172.2
2001	175.1	175.8	176.2	176.9	177.7	178.0	177.5	177.5	178.3	177.7	177.4	176.7	177.1
2002	177.1	177.8	178.8	179.8	179.8	179.9	180.1	180.7	181.0	181.3	181.3	180.9	179.9

Figure 3

If you need to look up an individual rate, such as the rate for May of 1996, this table supports this need extremely well. If, however, you wish to see how the rate changed in 1996 during the course of the year or to compare pattern of change in 1996 to the pattern in 1997, a graph would work much better, as you can see in figure 4.

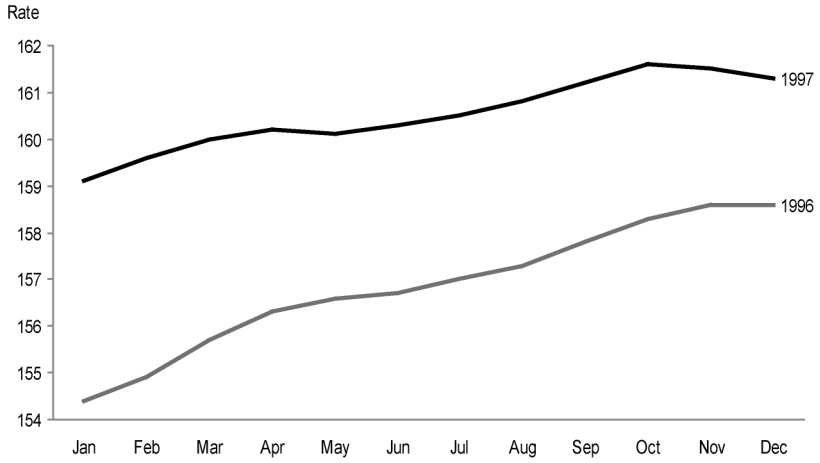


Figure 4

QUANTITATIVE VERSUS CATEGORICAL DATA

Quantitative information consists not only of numbers, but also of data that identifies what the numbers mean. It consists of quantitative data – the numbers – and categorical data – the labels that tell us what the numbers measure. The graph in figure 6 highlights this distinction by displaying the categorical data labels in green and the quantitative data labels in red.

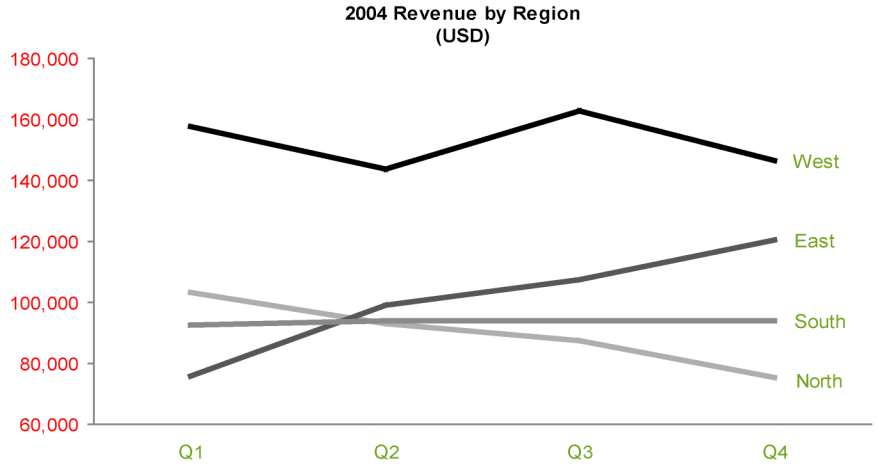


Figure 5

Figure 5 contains a quantitative scale along the vertical axis and a categorical scale along the horizontal axis. Most two-dimensional graphs consist of one quantitative scale and one categorical scale along the axes, although a familiar exception is the scatter plot, which has two quantitative scales.

Three Types of Categorical Scales

When used in graphs, categorical scales come in three fundamental types: nominal, ordinal, and interval. Nominal scales consist of discrete items that belong to a common category, but really don't relate to one another in any particular way. They differ in name only (that is, nominally). Items on a nominal scale, in and of themselves, have no particular order and do not represent quantitative values. Typical examples include regions (for example, The Americas, Asia, and Europe) and departments (for example, Sales, Marketing, and Finance).

Unlike a nominal scale, the items on an ordinal scale have an intrinsic order, but like a nominal scale, the items in and of themselves do not represent quantitative values. Typical examples involve rankings, such as "A, B, and C", "small, medium, and large", and "poor, below average, average, above average, and excellent".

Interval scales also consist of items that have an intrinsic order, but in this case they represent quantitative values as well. An interval scale begins its life as a quantitative scale, but is then converted into a categorical scale by subdividing the full range of values into a sequential series of smaller ranges of equal size, each with its own categorical label. Consider the quantitative range that appears along the vertical scale in figure 5 above. This range, from 55 to 80, could be converted into a categorical scale consisting of the following smaller ranges: (1) > 55 and <= 60, (2) > 60 and <= 65, (3) > 65 and <= 70, (4) > 70 and <= 75, and (5) > 75 and <= 80.

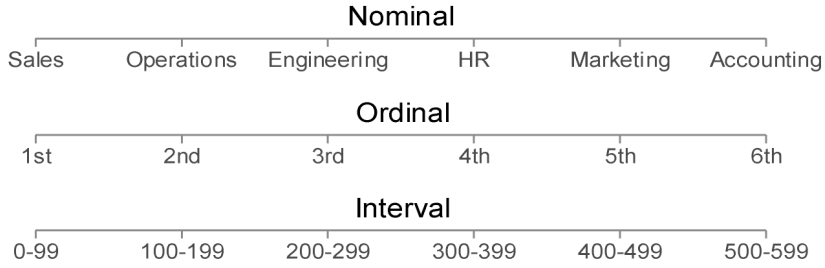


Figure 6

Here's a quick (and somewhat sneaky) test to see how well you've grasped these concepts. Can you identify the type of categorical scale that appears in figure 7?

Figure 7

Months of the year obviously have an intrinsic order, which begs the question: "Do the items correspond to quantitative values?" In fact, they do. Units of time such as years, quarters, months, weeks, days, hours, and so on are measures of quantity, and the individual items in any given unit of measure (e.g., years) represent equal intervals. Actually, months aren't exactly equal and even years vary in size occasionally due to leap years, but for most purposes of reporting and analysis, they are close enough in size to constitute an interval scale.

The relevance of this distinction between these three types of categorical scales will begin to become clear in the next section.

THE SEVEN COMMON RELATIONSHIPS IN QUANTITATIVE BUSINESS DATA

A number, by itself, is not very interesting. It is interesting, however, when you see that same number in relation to other numbers. The first question that you should always ask about a number is "Compared to what?" Numbers become meaningful only when compare to related numbers. Knowing that quarter-to-date revenue is $273,893 in itself isn't very revealing, but knowing that this is 19% below your revenue target for the quarter brings it alive and prompts action.

I've found that most of the meaningful relationships in quantitative business data can be classified into seven types. Knowing these types is the first step to knowing how best to display them, because these types correspond to particular means of display. Let's examine them, one at a time.

Time-Series Relationships

When quantitative values are expressed as a series of measures taken across equal intervals of time, this relationship is called a time series. No relationship is more common in quantitative business data. Studies have indicated that approximately 75% of all business graphs display time series. Time can be divided into intervals of varying duration, including years, quarters, months, weeks, days, and hours. The graph in figure 8 is a typical example.

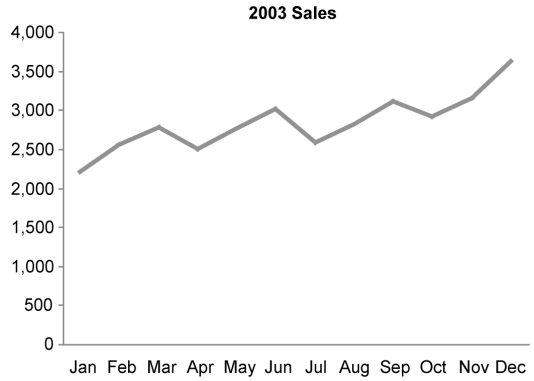


Figure 8

Seeing how values vary, moving up and down, through time, is very meaningful. Time series reveal trends and patterns that we must be aware of and understand to make informed decisions.

Ranking Relationships

When quantitative values are sequenced by size, from large to small or vice versa, this relationship is called a ranking. It is often meaningful in business to see things ranked, such as the performance of sales people or the expenses of departments. This not only reveals their sequence, but makes it much easier to compare values by placing those that are most similar near one another. Figure 9 shows a typical ranking relationship.

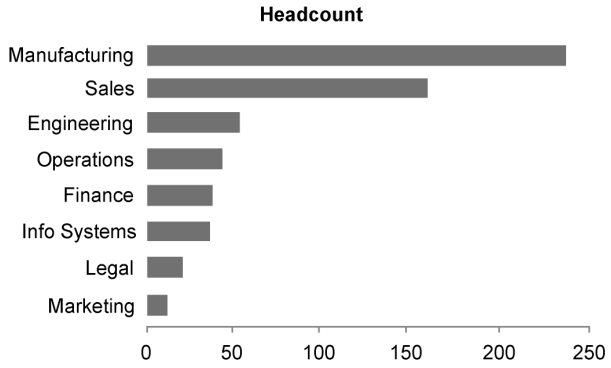


Figure 9

Part-to-Whole Relationships

When quantitative values are displayed to reveal the portion that each value represents to some whole, this is called a part-to-whole relationship. It is often useful to see how something is divided into parts, and the percentage of each part to the whole, such as how a market is divided up between competitors, or expenses are divided between regions, as shown in figure 10.

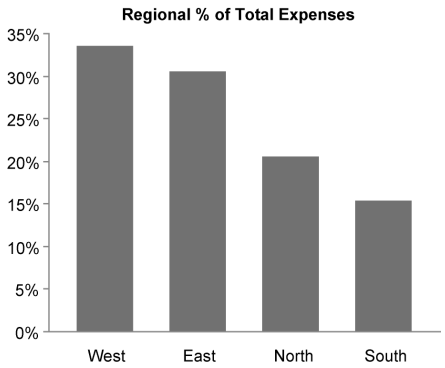


Figure 10

Deviation Relationships

When quantitative values are displayed to feature how one or more sets of values differ from some reference set of values, this is called a deviation relationship. The most common example in business is one that shows how some set of actuals (such as expenses) deviate from a predefined target (such as a budget), as shown in figure 11.

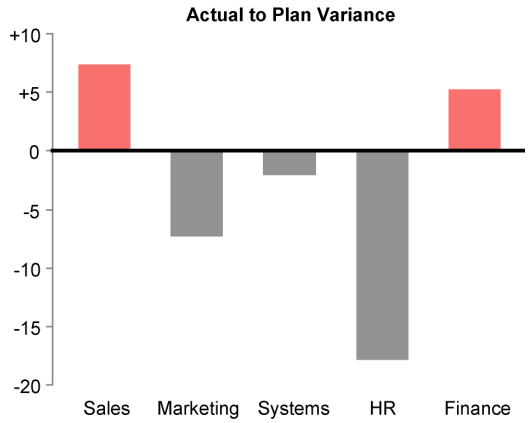


Figure 11

Distribution Relationships

When we show how a set of quantitative values are spread across their entire range, this relationship is called a *distribution*. We can often learn a great deal by examining the distribution of a set of values, especially the shape of that distribution, which reveals what's typical, if it is skewed in one direction or the other, and if there are gaps or concentrations. Figure 12 shows a distribution of values that is farily symmetrical, approaching what is called a *normal* or *bell-shaped curve*.

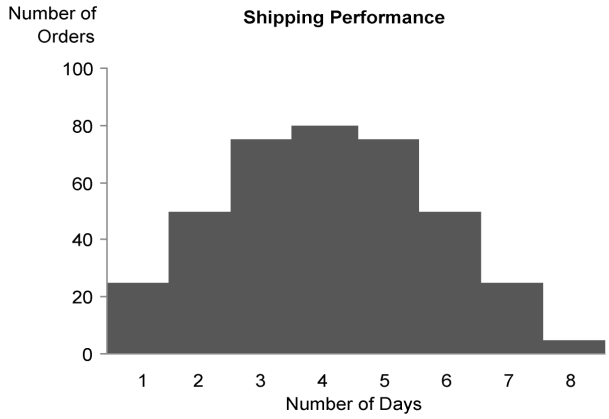


Figure 12

Correlation Relationships

When pairs of quantitative values, each measuring something different about an entity (for example a person, department, or product), are displayed to reveal if there is significant relationship between them (for instance, as one goes up the other goes up as well, or as one goes up the other goes down), this is called a correlation. Understanding correlations between quantitative variables can help us predict, take advantage of, or avoid particular behaviors. Figure 13 shows a correlation between employee's heights in inches (y axis) and their salary in dollars (x axis).

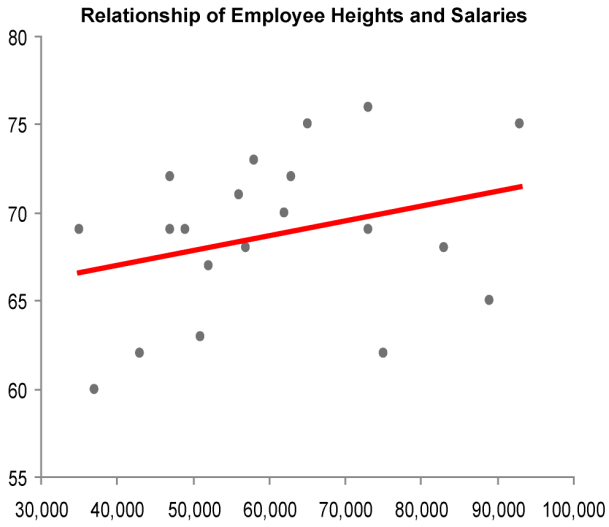


Figure 13

Nominal Comparison Relationships

I've saved the least interesting quantitative relationship for last. As you can see in figure 14, the categorical scale along the X axis is nominal. The four geographical regions do not relate to another in any particular order. In this case there is not particular relationship between the values. This is called a nominal comparison relationship. This graph provides a means to compare the regional values, but nothing more. It is always useful, whenever you prepare a graph that displays nothing but a nominal comparison, to ask yourself if another relationship could be featured that would make the graph more meaningful. In this case, simply arranging the regions in order of their quantitative values would produce a ranking relationship. Sometimes, however, discrete items in a categorical variable, like these geographical regions, need to be arranged in a particular order because people expect to see them arranged in that way.

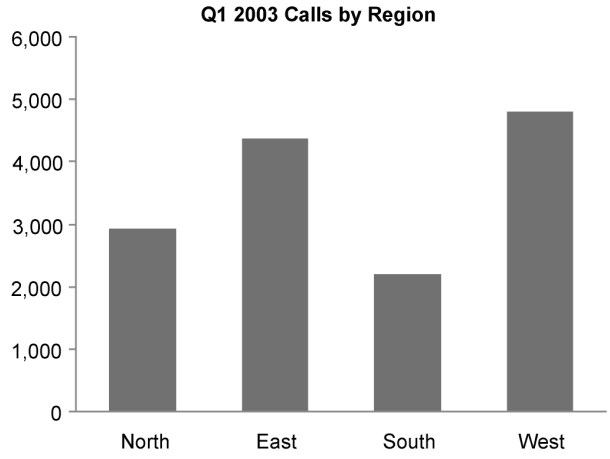


Figure 14

THE BEST MEANS TO ENCODE QUANTITATIVE DATA IN GRAPHS

Most graphs are two-dimensional, with one axis running vertically and the other running horizontally. Two-dimensional graphs work well because they use the two most powerful attributes of visual perception for encoding quantitative values: line length (or the length of shape similar to a line with an insignificant width, such as bars in a bar graph) and 2-D position (see figure 15).

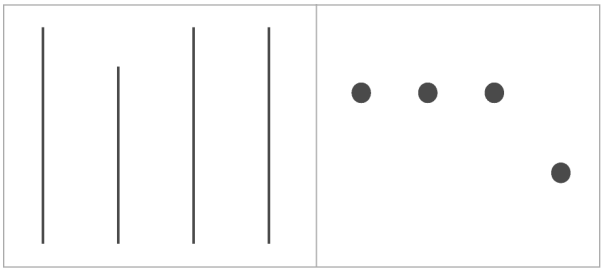


Figure 15

Only these two, of all the attributes of visual perception, including color, shape, and size, do an effective job of graphically representing quantitative values. "Can't the size of objects be used to encode quantitative values?" you ask. It can to a limited degree, in that you can tell that one of the circles in figure 16 is bigger than the other, but you can't easily determine how much bigger it is.

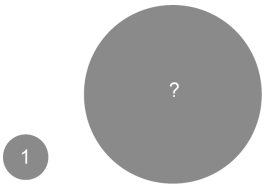


Figure 16

The larger circle is exactly 16 times the 2-D area of the smaller circle, but you probably guessed a different amount. This is because our ability to accurately compare 2-D areas is not well developed. It is best to avoid the use of 2-D areas to encode quantitative values in graphs, including the slices in a pie chart, whenever possible.

Four types of objects work best for encoding quantitative values in graphs: points, lines, bars, and boxes. Let's take a look at each in turn.

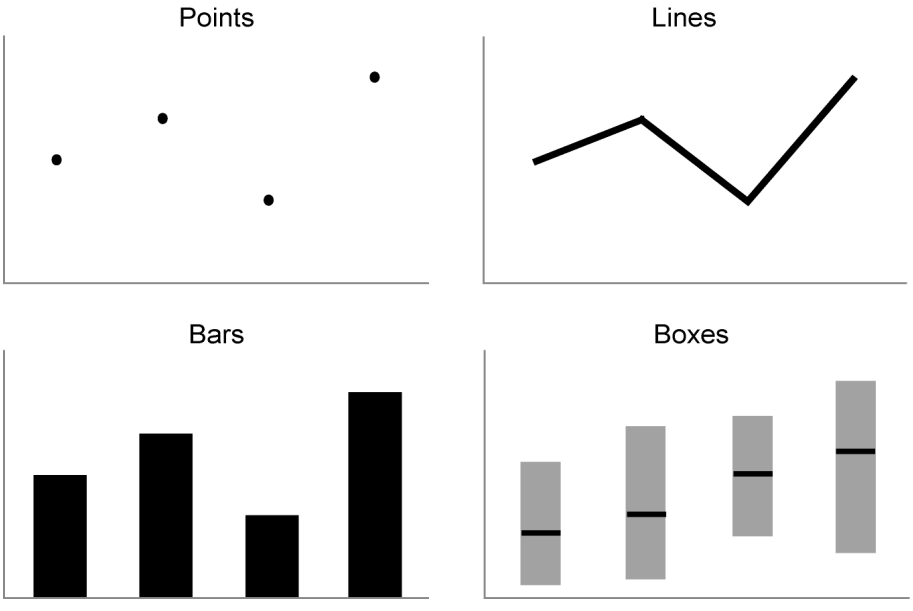


Figure 17

Points

Points are the smallest of the objects that are used to encode values in graphs. They can take the shape of dots, squares, triangles, Xs, dashes, and other simple objects. They have two primary strengths: (1) they can be used to encode quantitative values along two quantitative scales simultaneously, as in a scatter plot, and (2) they can be used to in place of bars when the quantitative scale does not begin at zero (Note: I'll explain more about this in the section on "bars" below.) Unlike lines, points emphasize individual values, rather than the shape of those values as they move up and down.

Lines

Lines connect the individual values in a series, emphasizing the shape of the data as it moves from value to value. As such, they are superb for showing the shape of data as it moves and changes through time. Trends, patterns, and exceptions stand out clearly.

You should only use lines to encode data along an interval scale. In nominal and ordinal scales, the individual items are not related closely enough to be linked with lines, so you should use bars or points instead. Lines suggest change from one item to the next, but change isn't happening if the items aren't closely related as sequential subdivisions of a continuous range of values. For instance, it is appropriate to use lines to display change from one day to the next or from one price range to the next, but not from one sales region to the next, as illustrated in figure 18.

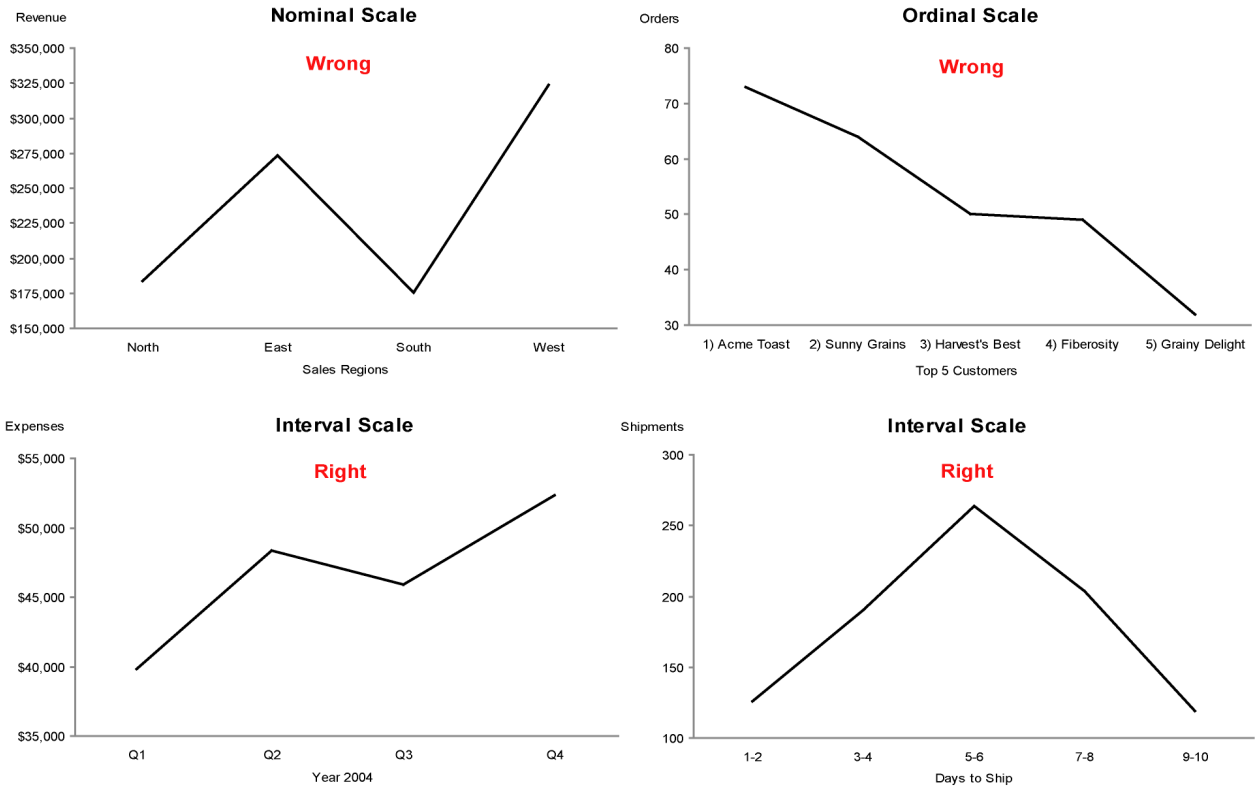


Figure 18

With interval scales, you are not forced in all cases to use lines; you can use bars and points as well. If you want to emphasize the overall shape of the data or changes from one item to the next, lines work best. If, however, you want to emphasize individual items, such as individual months, or to support discrete comparisons of multiple values at the same location along the interval scale, such as revenues and expenses for individual months, then bars or points work best.

Bars

Bars encode data in a way that emphasizes individual values powerfully. This ability is due in part to the fact that bars encode quantitative values in two ways: (1) the 2-D position of the bar's endpoint in relation to the quantitative scale, and (2) the length of the bar.

You probably recognize that these two characteristics correspond precisely to the two visual attributes that can be used to encode data in graphs. When you want to draw focus to individual values or to support the comparison of individual values to one another (see figure 19), bars are an ideal choice. They don't, however, do as well as lines in revealing the overall shape of the data. Bars may be oriented vertically or horizontally.

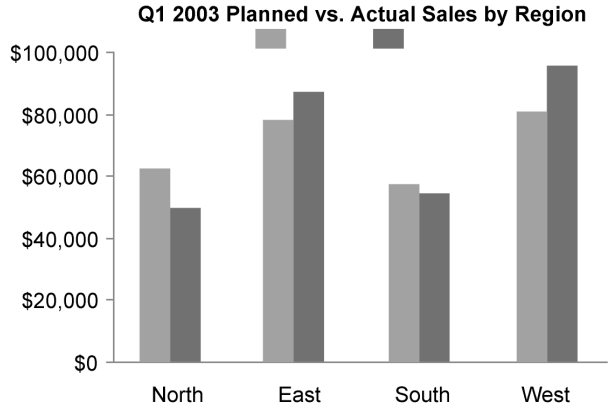


Figure 19

Whenever you use bars, your quantitative scale must include zero. This is because the lengths of the bars encode their values, but won't do so accurately if those values don't begin at zero. Notice what happens when you narrow the quantitative scale and use bars, as shown in figure 20. Actual sales in the North appear to be half of planned sales, but in fact they are 90% of the plan.

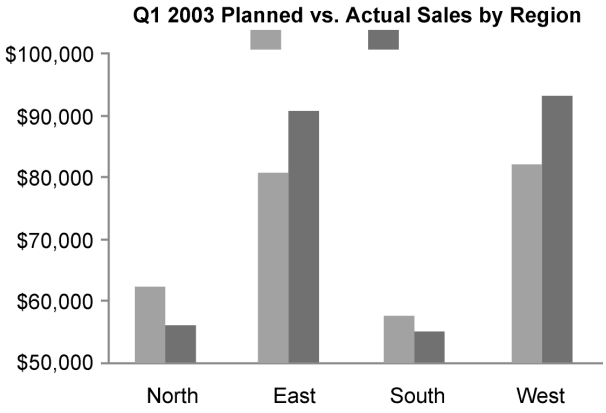


Figure 20

When you would normally use bars, but wish to narrow the quantitative scale to show differences between the values in greater detail, you should switch from bars to points, because points encode values merely as 2-D location in relation to the quantitative scale, which eliminates the need to begin the scale at zero.

Boxes

Boxes are a lot like bars, except that both ends encode quantitative values. When bars are used in this way, they are sometimes called range bars. They are used to encode a range of values, usually from the highest to the lowest, rather than a single value. In the 1970s a fellow named John Tukey invented a method of using rectangles (bars with or without fill colors) in combination with individual data points (often a short line) and thin bars to encode several facts about a distribution of

values, including the median (middle value), middle 50%, etc. He called his invention a box plot (a.k.a. box-and-whisker plot). Tukey's versions were designed for statisticians to use and required a little time to learn to read, but anyone can learn to read simpler versions like the one in figure 21 fairly easily.

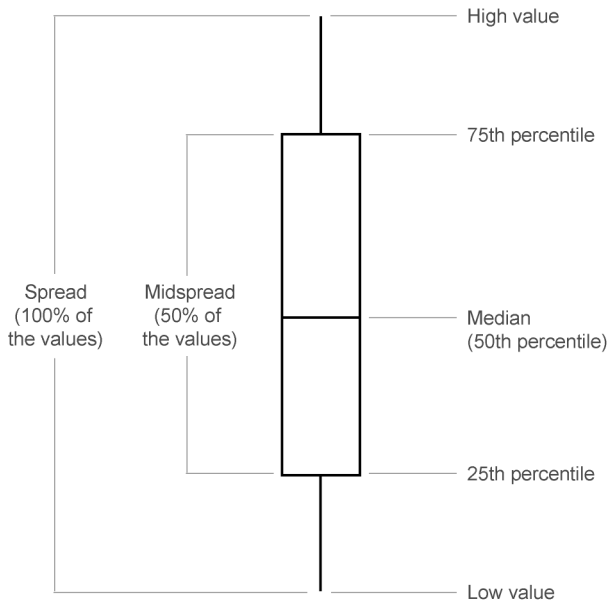


Figure 21

THE BEST PRACTICES FOR FORMATTING GRAPHS TO REMOVE DISTRACTIONS

Essentially, anything that doesn't contribute in an essential way to the meaning of a graph is a distraction that harms communication. The solution is simple: remove it. That snazzy map of the world that you'd like to place in the background of the plot area might look interesting, but it doesn't add any value. Grid lines might appear automatically when you create a new graph, but if they aren't needed to make sense of the data, they simply make it more difficult to focus on the data.

Some visual content in a graph doesn't represent actual data, but it serves a necessary role in supporting the data. The axis lines are an example of this. The lines themselves (not counting the tick marks) don't themselves encode data, but they do delineate the plot area of the graph, helping our eyes focus on the data. Items like these, which cannot be eliminated without losing something useful, should be visually subdued in relation to the data (for example, rendered as light gray rather than dark black)—just visible enough to do their jobs, and no more.

One of the worse distractions in graphs involves the misuse of color. A jumble of bright colors can visually overwhelm the viewer. Colors that are different for no reason, such as a different color per bar in a simple bar graph that contains a single set of values tempts our brains to search for a meaning for the differences. It is a good basic practice to use relatively soft colors in graphs, such as lowly saturated, natural colors found in nature, reserving the use of bright, dark, and highly saturated colors for those occasions when you need to make something stand out.

A STEP-BY-STEP GRAPH SELECTION AND DESIGN PROCESS

The steps that we'll cover in this section are presented as a logical sequence of design decisions, but the precise order of steps need not be cast in stone. In fact, once you've gone through the process a few times, it will become ingrained as a natural and only semi-conscious approach to graph design.

DETERMINE YOUR MESSAGE AND IDENTIFY YOUR DATA

Despite how obvious it might seem, I'm compelled to point out that the essential first step in graph design is determining what you want to say—compelled because this step is so often missed. It is not enough to simply take data that you've been tasked with presenting and turn it into a graph. When you speak to someone, you sift through content, then choose your

words and arrange them in a way that suits a particular communication objective. If you just spout everything that comes into your head at the moment however you feel like saying it, people will question your sanity—and for good reason.

Before you can communicate data, you must know what the data means and know what's important based on the needs of your audience. Only then can you trim away what's not pertinent, choose the best medium of display, and highlight what's most important.

DETERMINE IF A TABLE, GRAPH, OR BOTH IS NEEDED TO COMMUNICATE YOUR MESSAGE

Will the data be used to look up and compare individual values, or will the data need to be precise? If so, you should display it in a table.

Is the message contained in the shape of the data—in trends, patterns, exceptions, or comparisons that involve more than a few values? If so, you should display it in a graph.

If you need to do some of both, then display the data in both ways: in a table and in a graph.

If only a table is needed, then you must make a series of design decisions that I'm not addressing in this document. For guidance on this topic and a great deal more detail about graph design, you can consult my book titled *Show Me the Numbers: Designing Tables and Graphs to Enlighten*, Oakland, CA: Analytics Press, 2004.

DETERMINE THE BEST MEANS TO ENCODE THE VALUES

Here's where it comes in handy knowing the seven common relationships in quantitative business data. Knowing which one or more of these relationships best supports your message will help you narrow your choices regarding the right type of graph. The following table should give you what you need.

Relationship Type	Potential Encoding Methods
Nominal Comparison	--Bars (horizontal or vertical) --Points (if the quantitative scale does not include zero)
Time-Series	--Lines to emphasize the overall shape of the data --Bars to emphasize and support comparisons between individual values --Points connected by lines to slightly emphasize individual values while still highlighting the overall shape of the data
Ranking	--Bars (horizontal or vertical) --Points (if the quantitative scale does not include zero)
Part-to-Whole	--Bars (horizontal or vertical) *Note: Pie charts are commonly used to display part-to-whole relationships, but they don't work nearly as well as bar graphs because it is much harder to compare the sizes of slices than the length of bars.* --Use stacked bars only when you must display measures of the whole as well as the parts
Deviation	--Lines to emphasize the overall shape of the data (only when displaying deviation and time-series relationships together) --Points connected by lines to slightly emphasize individual data points while also highlighting the overall shape (only when displaying deviation and time-series relationships together)
Frequency Distribution	--Bars (vertical only) to emphasize individual values *Note: This kind of graph is called a histogram* --Lines to emphasize the overall shape of the data *Note: This kind of graph is called a frequency polygon.*
Correlation	--Points and a trend line in the form of a scatter plot

DETERMINE WHERE TO DISPLAY EACH VARIABLE

If the graph displays a single categorical variable, this step is simple—you simply associate it with one of the axes. With line graphs and any graph that involves an interval scale you should always position the categorical scale on the X axis (horizontal) and the quantitative scale on the Y axis (vertical). This is especially important when the interval scale involves a time series, because it is intuitive to display time horizontally, moving from left to right, rather than vertically.

If you are using bars to encode the data, and a time series is not involved, you have a choice between orienting the bars vertically, with the categorical scale on the X axis, or horizontally, with the categorical scale on the Y axis. Horizontal bars work especially well when either of these two conditions exist:

- The text labels associated with the bars are long
- There are many bars.

In both of these situations, the use of vertical bars would force you to attempt squeezing the labels along the horizontal axis, which would be difficult. Using horizontal bars solves this problem quite nicely, as shown in figure 22. Notice how easily they stack one on top of another.

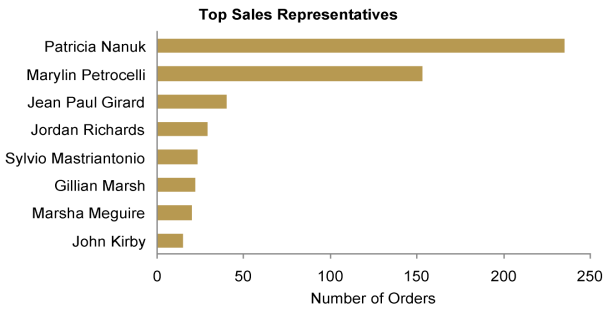


Figure 22

If the graph involves three variables, you must decide which to display along the axes and which to encode using distinct versions of another visual attribute, such as color. With a line graph, place the variable that is most important to your message along the X axis. With a bar graph, encode the variable whose items you want to make it easiest to compare using a method other than association with an axis, as shown in figure 23. Notice how much easier it is to compare bookings and billings than the regions, because they are positioned next to one another. It is usually best to encode the third variable using distinct colors, rather than any of the other available methods, such as different line or fill patterns. Just be careful to use colors that are still distinct, even when photocopied. This will guarantee that even those who are color blind will be able to see the distinctions.

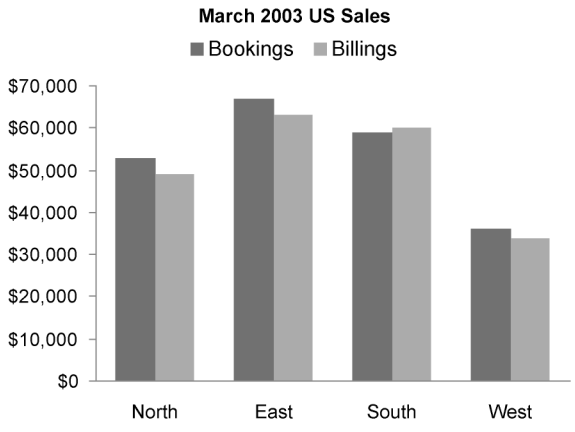


Figure 23

Most often, only a single graph is required to communicate a single quantitative message, but there are times when the number of variables that must be displayed will not fit into a single graph. Each of the two axes in a graph can be used to display a single variable, totaling two. You can also include another variable by including multiple lines, sets of bars, or sets of points that are each visually encoded in a distinct way, such as through the use of different colors. This brings the total up to three variables. What do you do if you need to display four variables? You could use a 3-D graph, adding a third axis (the z axis), but I recommend that you avoid this approach, because they are simply too hard to read. There is a solution, however, that works quite well. It involves what data visualization expert Edward Tufte calls "small multiples."

This solution involves a series of small graphs, all arranged together in a way that can be seen simultaneously. Each graph is alike, including consistent scales, differing only in that each features a different item of a categorical variable. Figure 24 should clarify what I'm describing. In it you see three graphs arranged horizontally, which each display sales, both bookings and billings, by geographical region—that's three variables. Each graph varies according to a fourth variable, which is sales channel (direct, distributor, or reseller). Keeping the quantitative scale consistent makes it is easy to compare the three sales channels.

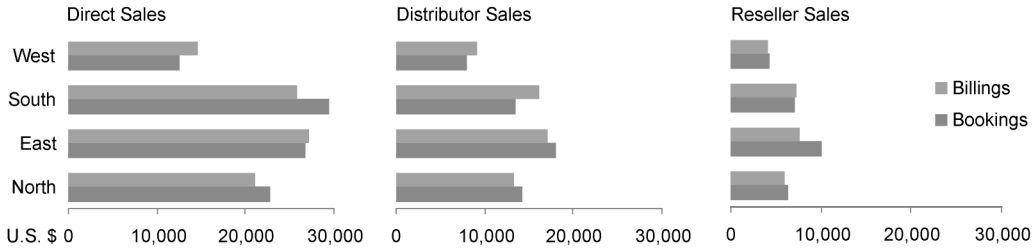


Figure 24

Using small multiples to support an additional variable is a powerful technique. Graphs can be arranged horizontally, vertically, or even in a matrix of columns and rows. If you need to display one more variable than you can fit into a single graph, select this approach.

DETERMINE THE BEST DESIGN FOR THE REMAINING OBJECTS

It's now time to make a series of design decisions that remain, including the scales and text. These decisions are concerned with the placement and visual appearance of items.

Determine the Range for the Quantitative Scale

Remember that if you are using bars to encode values, they must start from a value of zero on the quantitative scale, but if you are using lines, points, or a combination of line and points, you may want to narrow the scale. If the graph will be used for analysis purposes that require seeing the differences between values in as much detail as possible, narrowing the scale can be useful, as shown in figure 25. Generally, you should adjust the scale so that it extends a little below the lowest data value and a little above the highest.

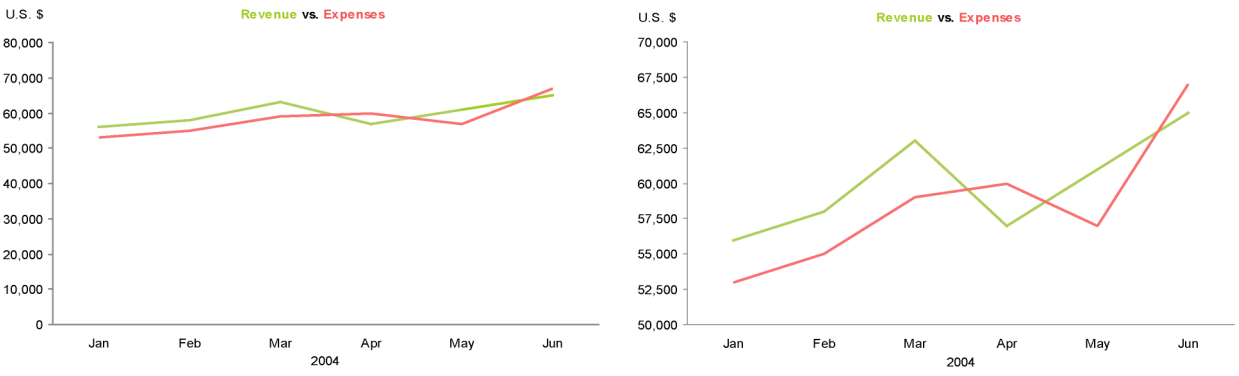


Figure 25

Be careful whenever you narrow the scale to make sure that it is obvious to your audience that you've done so and won't misread big differences between lines and points on the graph with big differences in their values, which might not be the case.

If you are using bars to encode the data, but your message could be better communicated by narrowing the scale, simply switch from bars to points. They aren't as visually prominent as bars and consequently don't emphasize individual values quite as forcefully, but points are a fine substitute for bars when you need to narrow the quantitative scale, as shown in figure 26.

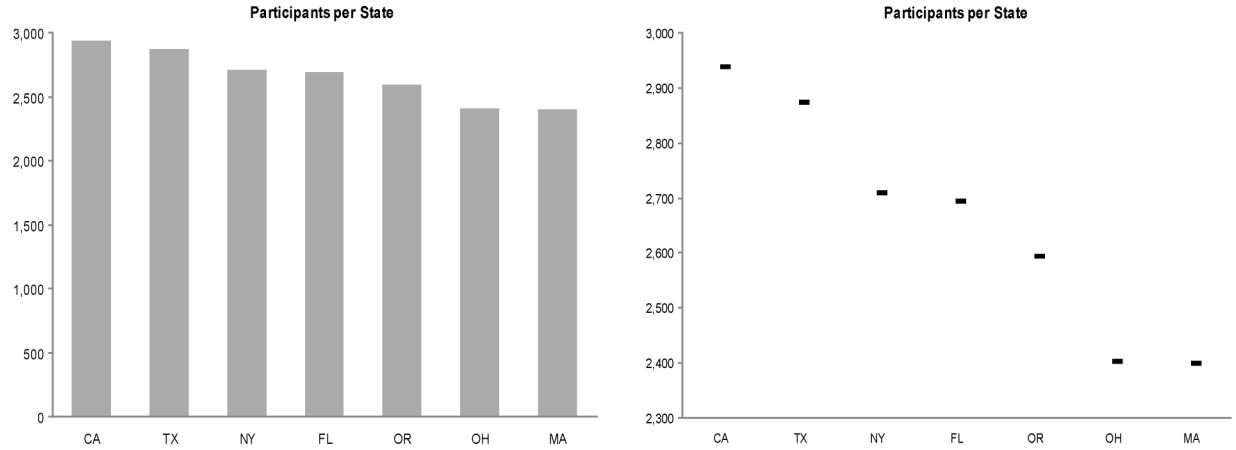


Figure 26

If a Legend Is Required, Determine Where to Place It

The more directly you can label data, the better. For instance in a line graph with multiple lines, if you can label the lines directly (for example, at the ends of the lines), the graph will be much easier to read. In a bar graph with multiple sets of bars, you usually need a legend, but you can make it much easier to read by arranging the labels to match the arrangement of the bars. Notice how much easier it is to tie the labels to the bars when they are arranged as shown in the right graph in figure 27, rather than the more usual way on the left. Notice also that the legend on the right doesn't have a border around it—it simply isn't necessary.

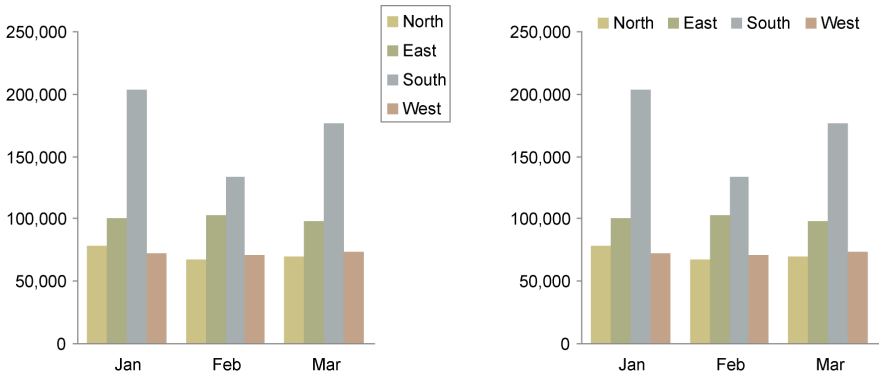


Figure 27

For Each Axis, Determine If Tick Marks Are Required and How Many

Tick marks are only necessary on quantitative scales, for they serve no real purpose on categorical scales. Even on quantitative scales, only major tick marks are necessary, with rare exceptions. A number between 5 and 10 tick marks usually does the job; too many clutters the graph and too few fail to give the level of detail needed to interpret the values.

Determine the Best Location for the Quantitative Scale

When the quantitative scale corresponds to the Y axis, it can be placed on the left side, right side, or on both sides of the graph. When it corresponds to the X axis, it can be placed on the top, bottom, or both. It is usually sufficient to place the quantitative scale in one place, but if the graph is so large that some values are positioned too far from the scale to adequately determine their values, placing the scale on both the left and the right, or the top and the bottom, will solve the problem.

When it only needs to appear in one place, the best choice of position depends on which values you want to emphasize or make easier to read. Placing the scale nearest to those values will accomplish this (see figure 28). Avoid placing the scale on the right side of the graph, however, unless really necessary to serve this purpose, because the scale so rarely appears only on the right that this might momentarily disoriented those who use the graph.

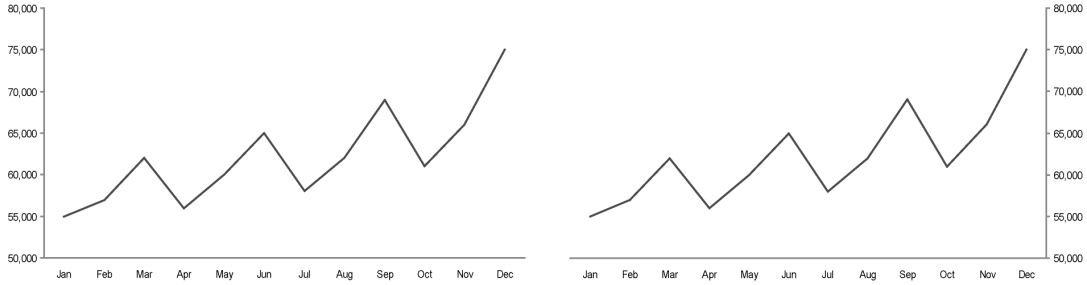


Figure 28

If the quantitative scale ranges between positive and negative values, the axis line should be positioned at zero, but the labels should be placed elsewhere so they won't interfere with the data. For instance, when the quantitative scale is on the X axis, it is usually best to place the text labels just below the plot area of the graph.

Determine If Grid Lines Are Required

Grid lines in graphs are mostly a vestige of the old days when graphs had to be drawn by hand on grid paper. Today, with computer-generated graphs, grid lines are only useful when one of the following conditions exists:

- Values cannot be interpreted with the necessary degree of accuracy
- Subset of points in multiple related scatter plots must be compared

Bear in mind that it is not the purpose of a graph to communicate data with a high degree of quantitative accuracy, which is handled better by a table. Graphs display patterns and relationships. If a bit more accuracy than can be easily discerned is necessary, however, you may include grid lines, but when you do, you should subdue them visually, making them just barely visible enough to do the job.

When you are using multiple related scatter plots and wish to make it easy for folks to compare the same subset of values in two or more graphs, a subtle matrix of vertical and horizontal grid lines neatly divides the graphs into sections, making it easy to isolate particular ranges of values, as shown in figure 29.

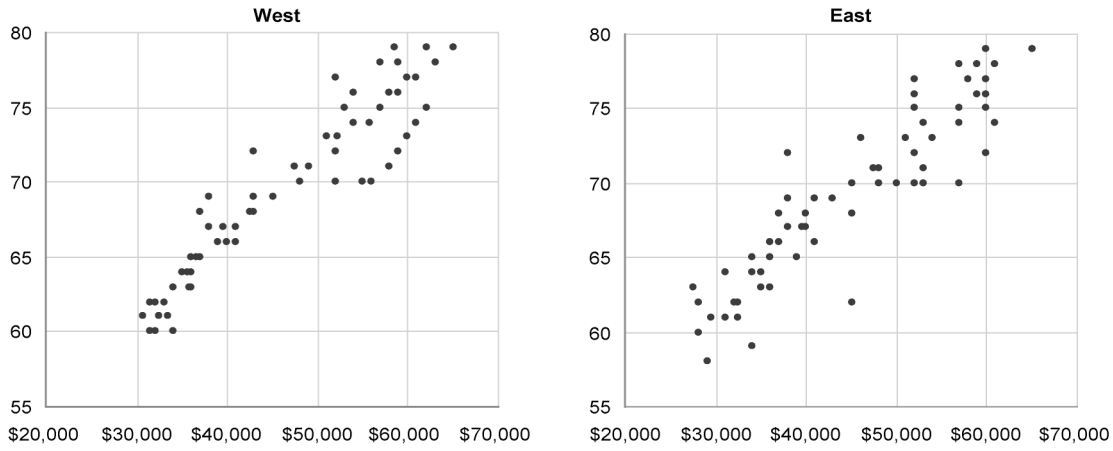


Figure 29

Determine What Descriptive Text Is Needed

Although the primary message of a graph is carried in the picture it provides, text is always required to some degree to clarify the meaning of that picture. Some text if often needed, including:

- A descriptive title
- Axis titles (unless the nature of the scale and its unit of measure are already clear)

Numbers in the form of text along quantitative scales are always necessary and legends often are, but we've already addressed these. Besides what I've already mentioned, it is often useful to include one or more notes to describe what is going on in the graph, what ought to be examined in particular, or how to read the graph, whenever these bits of important information are not otherwise obvious.

DETERMINE IF PARTICULAR DATA SHOULD BE FEATURED, AND IF SO, HOW

The final major stage in the process involves highlighting particular data if some data is more important than the rest. Perhaps you are showing how your company and all of your competitors compare in market share and you wish to highlight your company. Perhaps in a graph of revenue by month, the month of May needs to be featured because something special happened then. Whatever the reason, you have a number of possible ways to make selected data stand out.

One of the best and simplest ways is to encode those items using bright or dark colors, which will stand out clearly if you've used soft colors for everything else. Other methods include:

- When bars are used, place borders only around those bars that should be highlighted (see figure 30).
- When lines are used, make the lines that must stand out thicker.
- When points are used, make the featured points larger or include fill color in them alone.

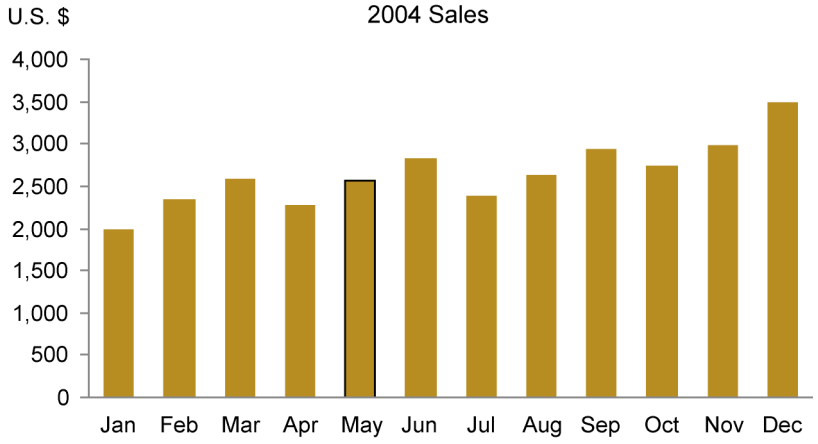


Figure 30

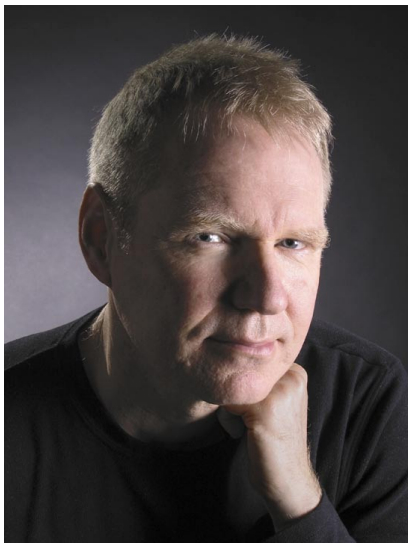
CONCLUSION

It is my hope that the concepts and practices in this white paper will become integrated into your graph designs without having to give it any real thought so that you can toss it aside or perhaps pass it on to someone else after using it for a short time. If you've read it and understand it, this is exactly what should happen.

Remember to follow this process for graph selection and design in order to communicate your information in the most effective manner:

- Determine your message and identify your data
- Determine if a table, graph, or combination of both is needed to communicate your message
- Determine the best means to encode the values
- Determine where to display each variable
- Determine the best design for the remaining objects
- Determine if particular data should be featured, and if so, how

Whenever you create a graph, you have a choice to make—to communicate or not. That's what it all comes down to. If you have something important to say, then say it clearly and accurately. These guidelines are designed to help you do just that.



ABOUT THE AUTHOR

Stephen Few has 24 years of experience as an IT innovator, consultant, and educator, specializing in business intelligence and information design. Today, as principal of the consultancy Perceptual Edge, he focuses on data visualization for the effective analysis and communication of quantitative business information. He writes the monthly data visualization column in DM Review, speaks frequently at conferences like those offered by The Data Warehousing Institute (TDWI) and DCI, and teaches in the MBA program at the University of California in Berkeley. He is also the author of the book *Show Me the Numbers: Designing Tables and Graphs to Enlighten*, and has a new book due out at the end of 2005 entitled *Information Dashboard Design: Beyond Gauges, Meters, and Traffic Lights*. More information about his current work can be found at www.perceptualedge.com.

Steps in Designing a Graph

Determine if a table, graph, or combination of both are needed

If...

- ...precision is required, or it will be used to look up or compare individual values — Create a table
- ...it will be used to see relationships, trends, patterns, or exceptions — Create a graph
- ...it will be used to see relationships — Create both

Determine the best means to encode the values

If your message features a...

- ...nominal comparison, ranking, or part-to-whole relationship — Use bars
- ..time-series, deviation, or distribution relationship, and...
  - ...your message is contained in the shape of the data (trends, patterns, or exceptions) — Use bars
  - ...you want to emphasize individual values or enable the comparison of individual values — Use lines
- ...correlation relationship — Use points

Determine where to display each variable

If your graph will include...

- ...one categorical variable, and...
  - ...any of the following are true:
    - the categorical scale is an interval scale
    - you are using lines to encode the data
    - you are using bars to encode the data and the the labels are not long or many

    — Place the categical variable on the X axis
  - ...you are using bars to encode the data — Place the categical variable on the Y
- ..two categorical variables — Determine the variable whose values you want to make easiest to compare, and then place the other variable on the X axis
- ..three categorical variables — Determine the variable whose values you least need to make easy to compare, and then create a series of "small multiples" and vary each graph by this variable. Follow the guidelines above for the remaining two variables.

Determine the best design for the remaining objects

- If your message requires that you zoom in for a closer view of differences between the values — Narrow the range of the quantitative scale, and if you are using bars, switch to points
- If a legend is required, and...
  - ...you are using lines — Label the lines directly, if possible
  - ...you are using bars — Place the legend above the plot area, with the labels arranged side-by-side in the same order as the bars
- Regarding tick marks — Use them only for quantitative scales and include between 5 and 10
- Regarding the location of the quantitative scale, if...
  - ...the graph can be read with the scale in one place only (left or right, top or bottom) — Place it in the position nearest the data that you wish to emphasize or make easiest to read
  - ...the graph is so large that it cannot be easily read with the scale in one position only — Place it in both positions (left and right or top and bottom)
- Regarding grid lines,
  - ...unless they are necessary to understand your message or to divide a scatter plot into sections for comparison — Leave them off, and whenever used, subdue them visually
- Regarding descriptive text — Include any text that is needed to clarify the meaning of your data and message

Determine if any data should be featured, and if so, how

- If particular data are especially important to your message — Highlight them using contrasting or more intense color, or by using one of the many other available visual attributes, such as a border or subtly-colored background