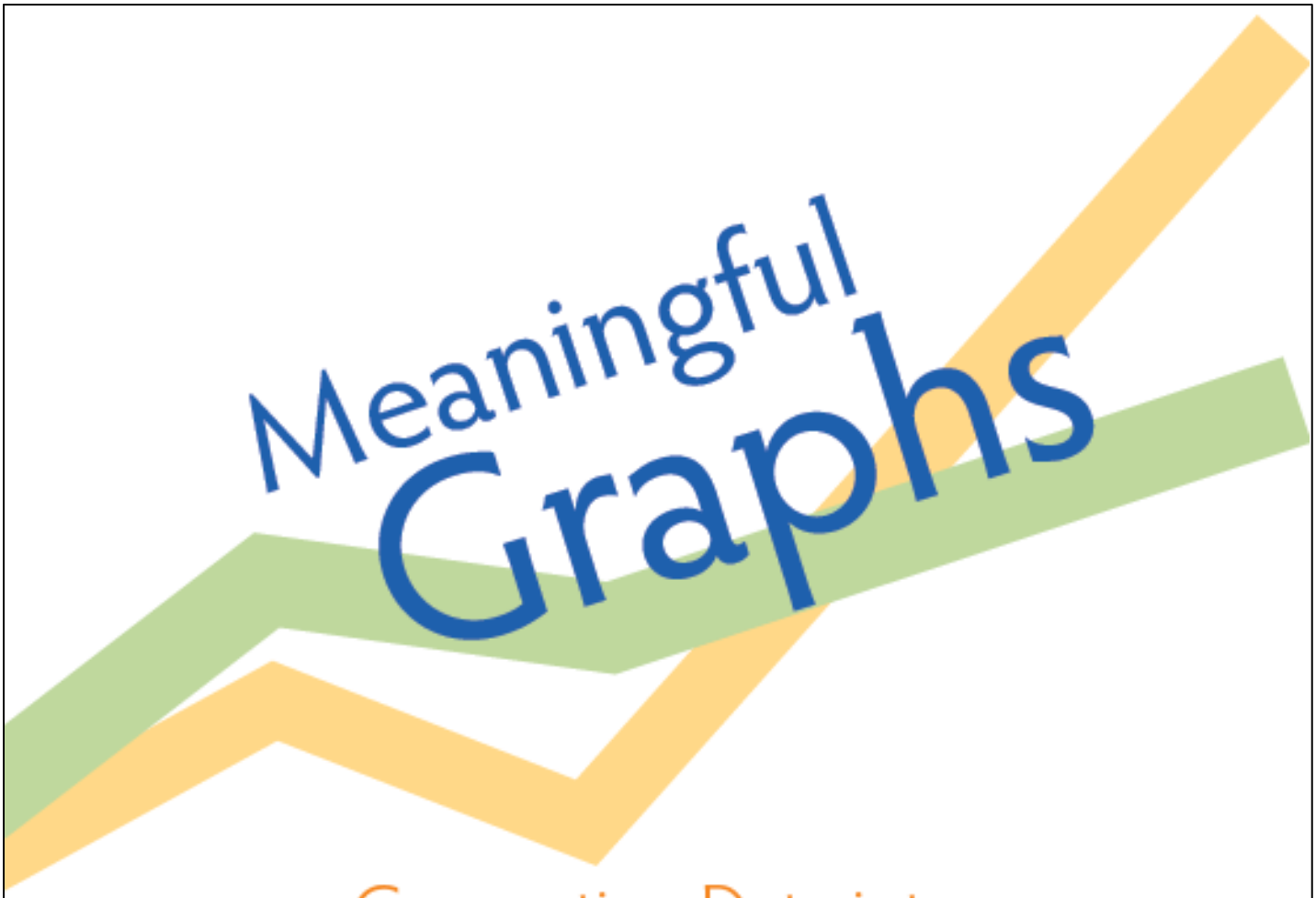

Meaningful Graphs

Converting Data into Informative Excel® Charts

James M. Smith, Ph.D.

First edition, 2014.

ISBN 978-0-9860549-0-7 (paperback)

Book design by DesignForBooks.com

Printed in the United States of America.

Contents

Appendices

Chartjunk

"Chartjunk" is a term coined by Tufte in his 1983 book *The Visual Display of Quantitative Information*. It refers to all elements of a chart that do not convey information. This includes pictures and icons, background graphics, template material, shading and color for no purpose, redundant data, and the like. In this book Tufte introduced the concept of data-ink, based on the principle that "a large share of ink on a data graphic should present data-information, the ink changing as the data change" (Tufte, 1983, p. 93). Comparing the data-ink to the total ink used to produce the graphic, yields a conceptual data-ink ratio.

$$\frac{\text{data-ink}}{\text{total ink used to print the graphic}}$$

The higher the ratio, the greater the proportion of total ink devoted to data and the less unnecessary material the graphic contains. According to Tufte (1983), the goals should be to "maximize the data-ink ratio, within reason" (p. 96) and to "erase non-data-ink, within reason" (p. 96). "Within reason" is a key phrase since at times the addition of non-data-ink may allow observers to read charts more quickly. For our purposes, suffice it to say that you should consider removing all *unnecessary* material from a chart.

If you want to see some classic examples of chartjunk, take a look at the USA Today Snapshot charts on page 1 of this daily newspaper. They are usually about an interesting topic, are simple and colorful, and are typically loaded with chartjunk (background pictures, totally unrelated content, etc.). Though I enjoy looking at these charts, I wouldn't want to incorporate any chartjunk similar to this in a professional presentation.

Chartjunk is not merely confined to over-the-top elements gratuitously added to a chart; some is much more subtle. What's the chartjunk in Fig. 3.5?

TIP

If you're unfamiliar with the *USA Today* Snapshot charts, Google "USA Today Snapshot Archive," and you'll find numerous examples.

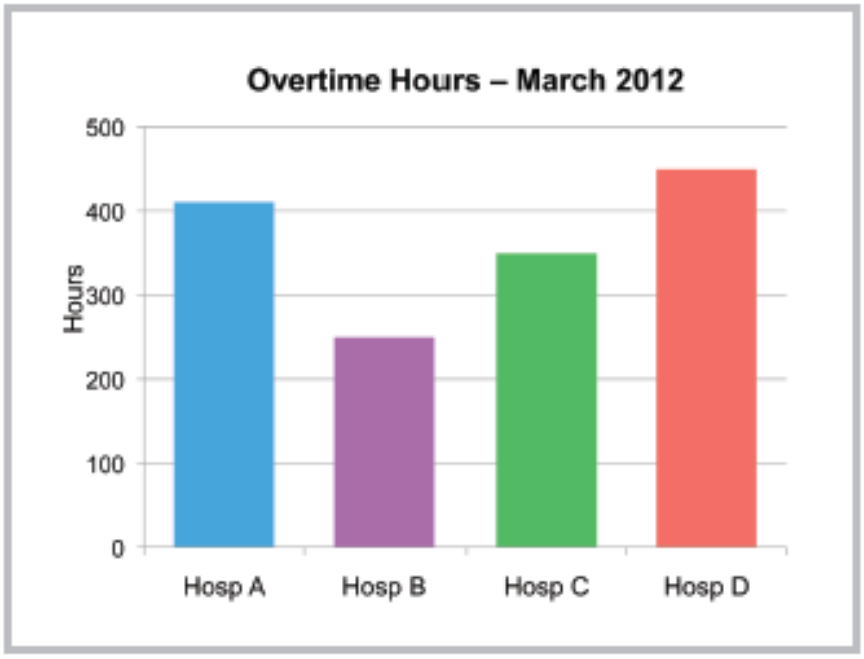

Fig. 3.5. Chartjunk—Illustration A

TIP

Sometimes, as we'll see later, in a single-series column or bar chart, one bar or column will be given a different color or a different shade of the same color to highlight it. That use of color is fine.

The different colors for each of the columns are chartjunk. Colored columns or different shades of the same color are used typically to denote different series of data (as in the Medicare discharge chart, Fig. 3.3), but in Fig. 3.5, in which there is only one series of data, the colors of the different columns have no meaning whatsoever. Even though they have no intentional meaning, sometimes colors have different psychological meanings (e.g., green = good, red = bad). I have no idea how the habit of coloring columns or bars different colors in a single series column or bar chart originated, but I see it used quite often.

What's the chartjunk in Fig. 3.6?

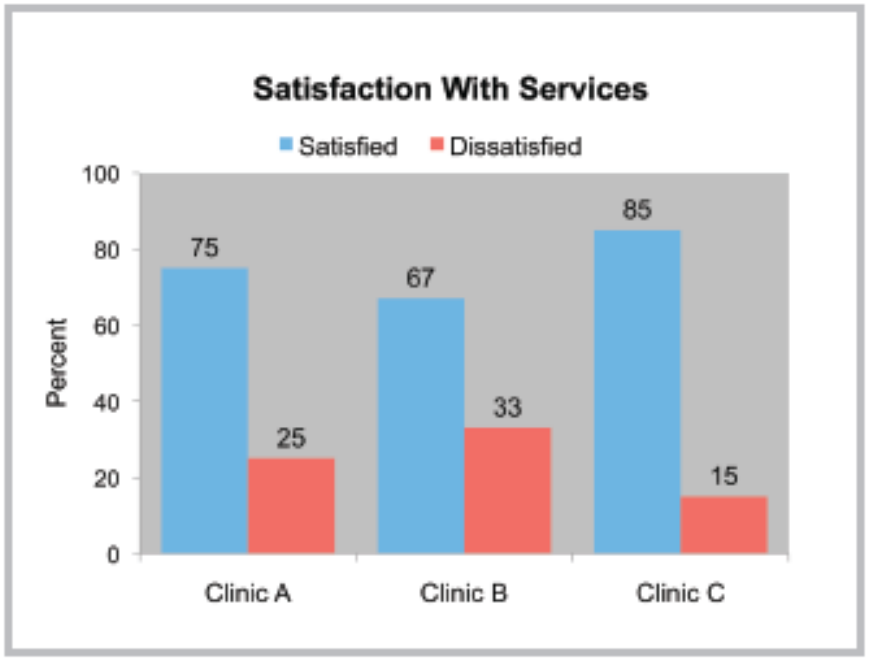

Fig. 3.6. Chartjunk—Illustration B

IN PRACTICE

On occasion I've seen charts with columns showing the frequencies for each response category (e.g., excellent, very good, good, fair, poor) for each of the x-axis categories. This turns a chart, which should be designed to convey information quickly and easily, into the equivalent of a tally sheet.

The plot area fill color (the background gray) is chartjunk. It has no meaning and sometimes makes a chart more difficult to read. But there's another, even more problematic, example of chartjunk in this chart.

Either the satisfied scores or the dissatisfied scores are chartjunk. Since the two scores for each clinic sum to 100%, if you know one, you automatically know the other. Presenting both values as part of the chart is unnecessary and distracting, forcing the reader to look at both series of data when one would suffice to convey the information. Whether you choose to present percent satisfied or percent dissatisfied is up to you.

What's the chartjunk in Fig. 3.7?

End of this section excerpt

Note how the category labels are much easier to read both as a result of their horizontal orientation and the ability to use two lines to accommodate the longer labels. In addition, there is much more room allotted to the chart plot area. The chart plot area in Fig. 4.15 is over 80% larger than the chart plot area in Fig. 4.14.

One area where bar charts should *not* be used is to show trends over time, even for short intervals. Most observers expect to see time conveyed from left to right. Asking observers to view a chart with time moving in a vertical direction (up or down) is perceptually awkward.

Dot Plots

Dot plots are used to present information on categorical data when the number of categories is large and, as a consequence, the use of multiple very thin columns or bars would appear cluttered (Robbins, 2005). Markers ("dots"), placed at the value for each category, take the place of columns or bars. In addition to simplifying the appearance of the chart, another advantage of dot plots is that the value axis can be truncated, since columns and bars are not involved.

Dot plots are usually portrayed in a horizontal (bar chart) format to enable the labeling of the various categories. Dot plots also frequently employ a Pareto format from highest to lowest or vice versa depending on the data. Figure 4.16 presents a dot plot of falls per 1,000 patient days for 14 nursing homes with the Burke nursing home highlighted.

Creating a dot plot with the category names on the vertical axis is not an option in Excel, but can be accomplished with some add-ons to Excel. However, if you have only an occasional need for a dot plot, it can easily be created in Excel if you are willing to do some hand work. The method that produced the chart in Fig. 4.16 is described in Appendix C.

Although dot charts can be used to show static differences among categorical elements, they present no data-over-time context for any of the categories in the chart. A small multiples (multipanel) display

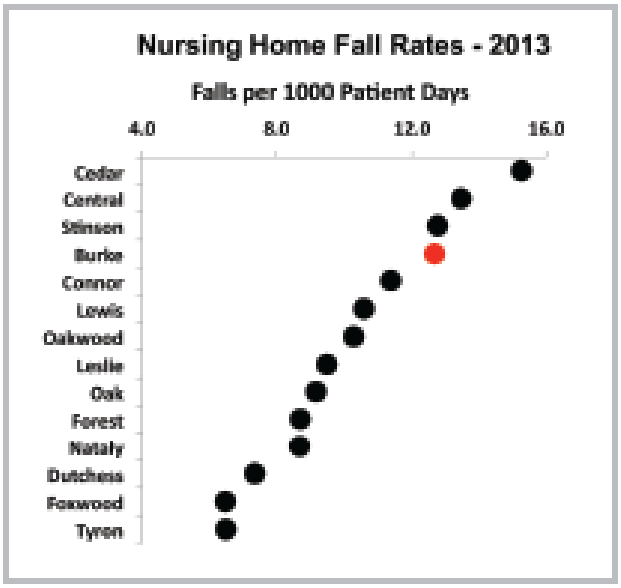

Fig. 4.16. Dot Plot With Category Labels on the Vertical Axis

of fall rates for each of the nursing homes over the 12 months of the year (using the same value axis range for each chart) would enable a more thorough exploration of differences in fall rates.

Balestracci (2009, pp. 131–135) uses a very dramatic illustration of the problem associated with using only aggregate summary statistics. He presents tabular data on the means and standard deviations of mortality rates for three hospitals over the course of 30 months. All three have similar means and standard deviations. Yet, despite the comparability of means and standard deviations, one hospital has a significant increasing trend over the course of the 30 months, one a significant decreasing trend, and one a stable but highly variable trend. Clearly, the aggregate data misrepresented the situation and would have led to an erroneous conclusion, namely, that all three hospitals had comparable and presumably stable mortality rates.

If the aim of the chart is to indicate where a particular facility appears among its peers *and the specific names of the peers are not important*, a simple line chart with markers may be used with the line deleted, as in Fig. 4.17.

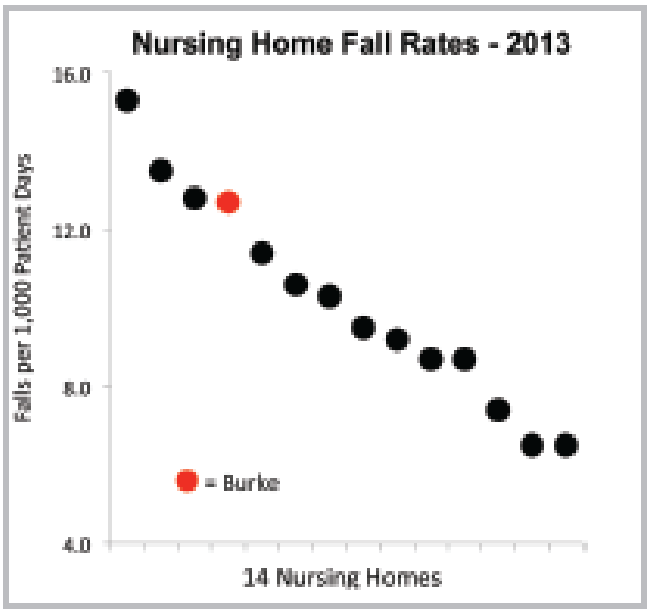

Fig. 4.17. Dot Plot Using a Line With Markers (Line Deleted)

TIP

If markers are too close together to select one over others nearby, open the datasheet, temporarily change the value for the data point of interest to separate it from the others in the chart, click on it in its changed position, change the format as desired, and reenter the correct value for this data point to place it back where it belongs.

HOW-TO. Figure 4.17 is a line chart with markers with the data sorted in descending order. The lines have been removed so that only the markers remain. Lines can be removed by double left clicking on the line | Format Data Series | Line Color | No Line.

The circle for Burke was highlighted by left clicking on the marker for Burke (all markers will be highlighted) and then left clicking again to highlight the marker for Burke only. Double left click on the highlighted Burke marker | Format Data Point | Marker Fill | Solid Fill | Red and Marker Line Color | Solid Line | Red.

The single legend for Burke was created by inserting a circle shape the same size as the markers in the chart, filling it with red color, and pairing it with a text box with the entry "= Burke."

End of this section excerpt

Unequal Time Intervals

Sometimes data are collected on a time-as-available basis, resulting in data at unequal time intervals, as in Table 6.2.

Table 6.2. Data Collected at Unequal Time Intervals

Date	Value
1/1/12	4
1/14/12	5
1/31/12	2
4/1/12	3
6/2/12	5

To show an accurate representation of a trend over time when data collection occurs at irregular intervals, the x-axis should be spaced as a calendar would (i.e., including x-axis intervals for periods for which no data were collected). Thus, the chart of the data in Table 6.2 should show February, March, and May on the x-axis, even though no data were collected during these months.

If unequal time intervals are displayed as equal time intervals (i.e., as text), a different and erroneous picture of the trend over time will emerge (Jelen, 2011; Robbins, 2005). In the data in Table 6.2, three of the five measures were obtained in the first month of this five-month period. The chart on the left of Fig. 6.2, with a date-based x-axis, correctly displays the early decline in this five-month period (placing the data points exactly where they belong, even mid-month). The chart on the right in Fig. 6.2, with a text-based x-axis (i.e., equally spaced x-axis intervals) gives the erroneous impression that the decline occurred in the middle of this five-month period.

If data are entered in the form of month/day/year, Excel will automatically display the appropriate date scale axis, spacing the categories as they would appear on a calendar. However, Jelen (2011)

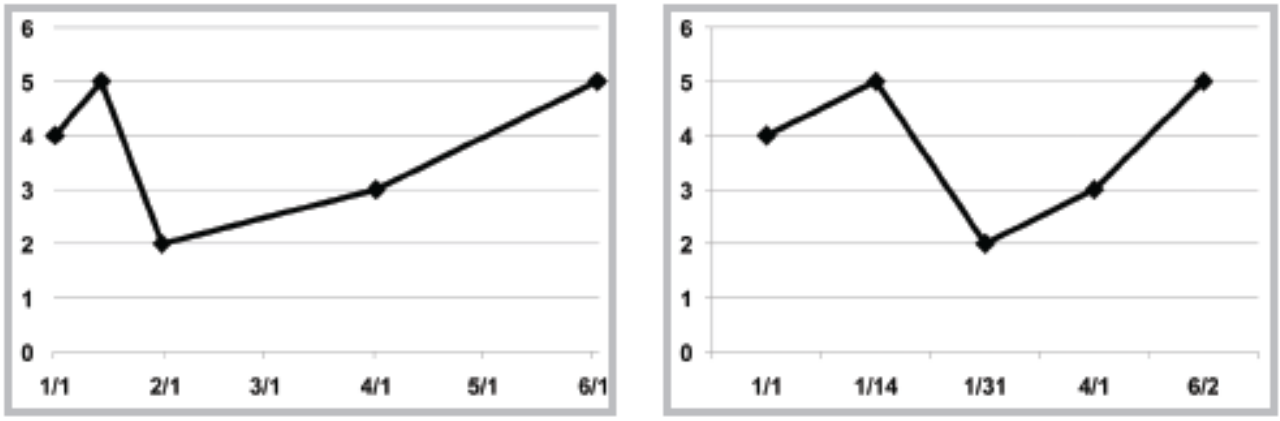

Fig. 6.2. Date-Spaced Scale Versus Equal-Spaced Scale

cautions that you should always check to be sure that this happens. He notes, for example, that if just numeric years are entered (e.g., 2008, 2009, 2011), Excel will consider this a text category, and the year 2010 will not appear on the x-axis.

HOW-TO. If you want to change equal time intervals (i.e., text format) to unequal time intervals (i.e., calendar format), double left click on the x-axis | Format Axis | Axis Type | Date Axis. Since a line drawn between irregularly spaced points suggests a specific pattern between these points, you might want to consider using just the markers without a connecting line.

Sometimes (e.g., in tracking performance over time) the information of interest might be the overall trend and the specific value of the last data point. In this case, create a line chart without data labels. Left click on the last data point (all data points will be highlighted); left click again on the last data point, and only that data point will be highlighted. Right click on this last data point | Add Data Label. The data label for the last data point will be added.

TIP

If you want to see how the equal and unequal time intervals work in practice, insert a line chart with markers into a PowerPoint slide. Change the category names to 2007, 2008, 2010, and 2011; leave the sample data alone. Expand the slide view | double left click on the x-axis | Format Axis | and in the Axis Type toggle back and forth between Text axis and Date axis. Note how the x-axis in the chart changes.

End of this section excerpt

daily basis to enable more rapid evaluation of a specific intervention.

Let's apply the run chart methodology to the data on door-to-needle time presented in Fig. 6.1. Figure 6.10 presents these data with the pre-intervention median added.

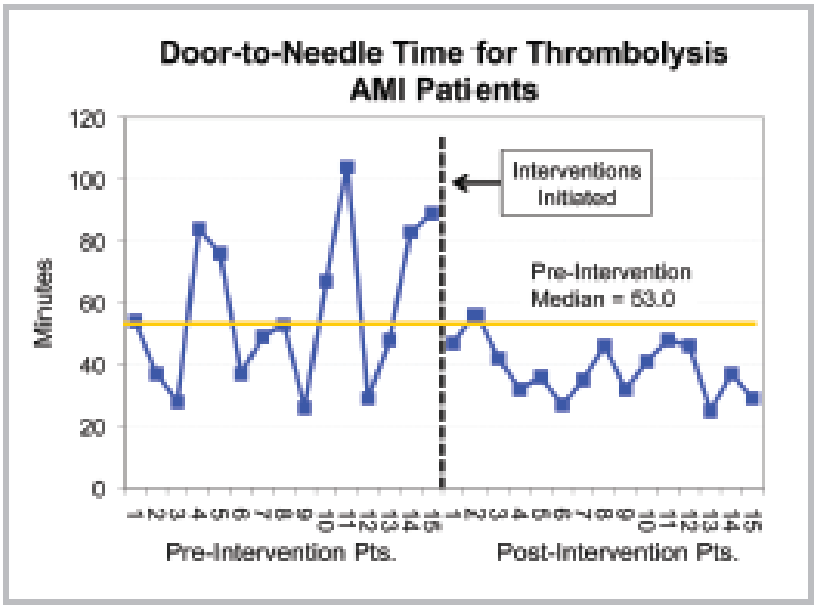

Fig. 6.10. Run Chart of Door-to-Needle Times for 30 Successively Admitted AMI Patients

TIP

Markers were added to the lines in Fig. 6.10 to facilitate counting. In this instance, it may not have been necessary since the lines show noticeable fluctuation from one point to the next but in instances where several values in a row are similar and the x-axis categories are close together, markers would be helpful.

When the door-to-needle times for six consecutive patients after the intervention were below the median for the baseline period (post-intervention Patients 3–9), this was an indication that the interventions introduced a statistically significant change in the process. The post-intervention period also appears to be associated with reduced variability. These are two critical aims of process improvement: improvement in the overall level of functioning and a reduction in the variability of the process.

There are additional techniques that can be used to identify special causes in a line chart over time (most notably, control charts), but they are beyond the scope of this discussion. Several sources

provide information on the use and calculation of control charts with a focus on data from health care settings (Balestracci, 2009; Carey & Lloyd, 1995; Kelley, 1999).

Note that run charts (as well as control charts) can indicate whether your process has been affected by special causes, either positive or negative. If no special causes are identified, only common causes are present. These may result in a process that is unsatisfactory because either the average level of performance or its high degree of variability does not meet your expectations. In this case, the overall process should be examined further to identify any elements of the process which could be redesigned to improve performance. For example, if missed appointments are stable but at a level that is unsatisfactorily high, data on the process should be obtained to determine whether any elements of the appointment process could be modified to reduce the number of missed appointments.

Data Exploration

An important first step in exploratory data analysis is to examine your data unencumbered by hypotheses or summary statistical measures. Balestracci (2009) refers to this as "plotting the dots." Yau (2011) advises that you should "learn all you can about the data, and the visual storytelling will come natural" (p. 328). Cairo (2013), in relation to the creation of infographics, notes, "Sometimes it is not the story which leads you to search for a particular kind of data. Sometimes it is data that leads you to a story" (p. 160).

Here's a simple but compelling example of letting the data tell their own story. It was first brought to my attention by its description in Wainer (1997). It's a story of World War II and a man named Abraham Wald.

Abraham Wald was a brilliant mathematician who immigrated to the United States when the Nazis invaded Austria; he took a position at Columbia University in New York City. During World War II, the British were concerned about the losses of their bombers to enemy fire and wanted to put additional armor on the most

vulnerable parts of the aircraft. To determine these vulnerable areas, they collected data on the damage due to enemy fire by meticulously measuring the precise location of each of the bullet holes in their bombers when they returned from a mission. The UK Air Ministry provided Wald with these data and asked if he could conduct a mathematical vulnerability analysis that would indicate where they should place additional armor.

One of the things that Wald did was to indicate the location of the bullet holes on a schematic outline of a bomber. After a period of time doing this, the schematic outline of a bomber marked with the location of bullet holes looked like the outline on the left in Fig. 6.11. After a longer time, it looked like the outline on the right.

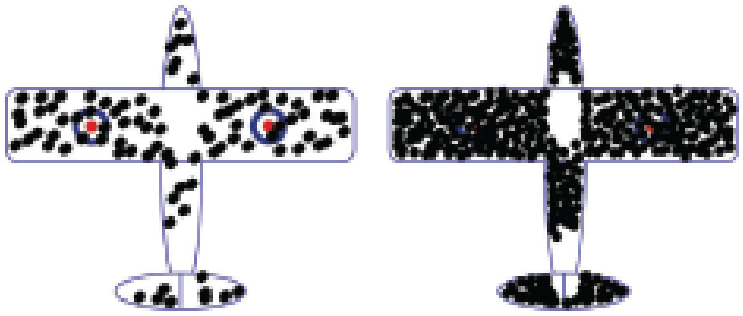

Fig. 6.11. Pattern of Bullet Holes in British Bombers

Where should they put the extra armor? Think about it.

Many might be inclined to place the extra armor on those areas that had sustained the most damage. However, Wald suggested that they put the extra armor in the areas where there was no damage. Why? Because if a bomber was hit in either of these two areas, it never made it back to the airfield to have its bullet holes measured—it was downed by enemy fire. Wald didn't have to wrench the solution from the data, he just let the data speak, and they pointed to the solution.

End of this section excerpt

Index