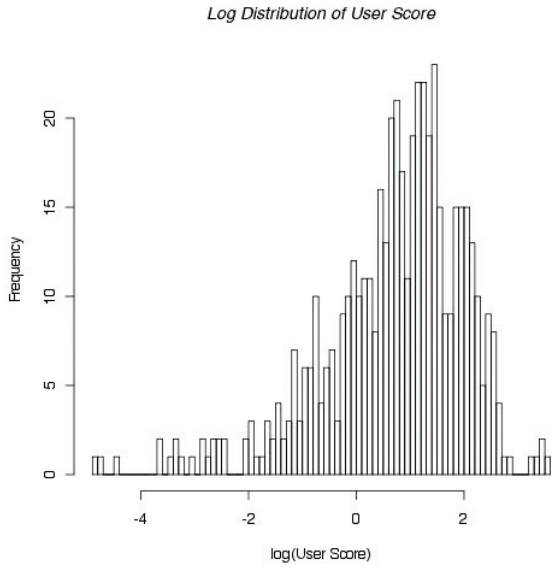


FIGURE 12-20 *Frequency distribution with log of user score*

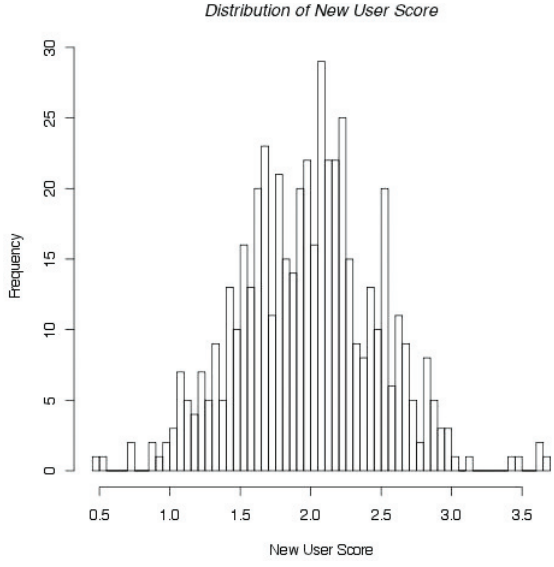


FIGURE 12-21 *Frequency distribution of new user scores*

Another idea may be to analyze the stability of price distributions over time to see if the prices offered to customers are stable or volatile. As shown in a graphic such as Figure 12-22, the prices appear to be stable. In this example, the user score of pricing remains within a tight band between two and three regardless of the time in days. In other words, the time in which a customer purchases a given product does not significantly influence the price she is willing to pay, as expressed by the user score, shown on the y-axis.

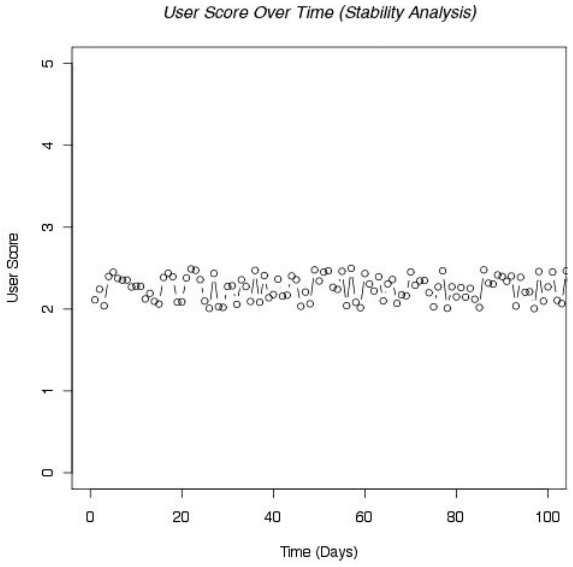


FIGURE 12-22 *Graph of stability analysis for pricing*

By this point the data scientist has learned the following about this example and made several observations about the data:

- Most user scores are between two and three in terms of their price sensitivity.
- After taking the log value of the user scores, a new user scoring index was created, which recentered the data values around the center of the distribution.
- The pricing scores appear to be stable over time, as the duration of the customer does not seem to have significant influence on the user pricing score. Instead, it appears to be relatively constant over time, within a small band of user scores.

At this point, the analysts may want to explore the range of price tiers offered to customers. Figures 12-22 and 12-23 demonstrate examples of the price tiering currently in place within the customer base.

Figure 12-23 shows the price distribution for a customer base. In this example, loyalty score and price are positively correlated; as the loyalty score increases, so do the prices that the customers are willing to pay. It may seem like a strange phenomenon that the most loyal customers in this example are willing to pay higher prices, but the reality is that customers who are very loyal tend to be less sensitive to price fluctuations or increases. The key, however, is to understand which customers are highly loyal so that appropriate pricing can be charged to the right groups of people.

Figure 12-24 shows a variation on 12-23. In this case, the new graphic portrays the same customer price tiers, but this time a rug representation (Chapter 3) has been added at the bottom to reflect the distribution of the data points.

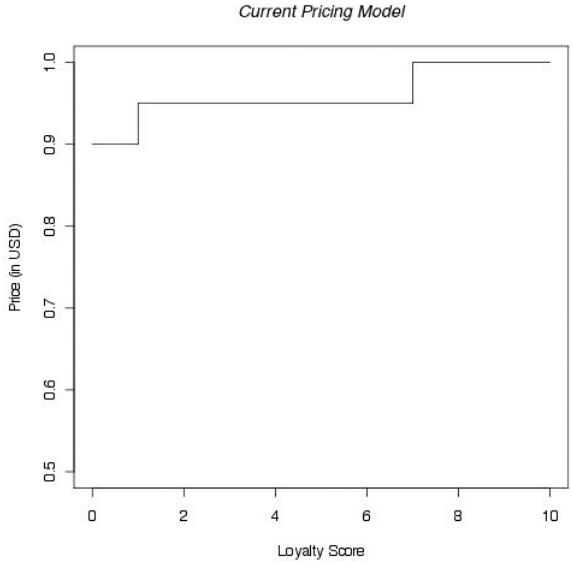


Figure 12-23 *Graph comparing the price in U.S. dollars with a customer loyalty score*

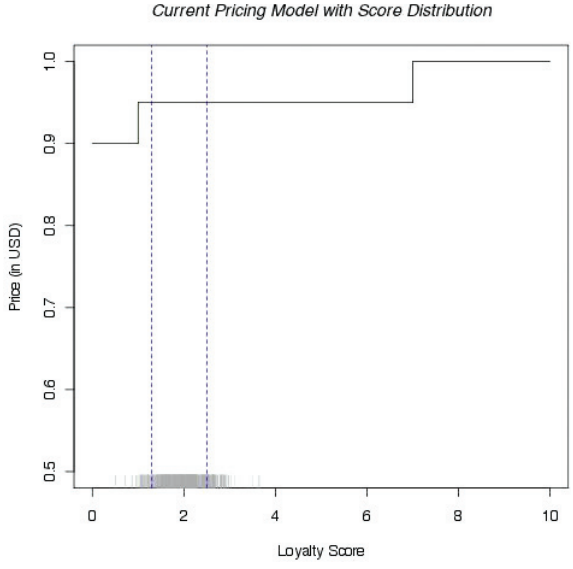


Figure 12-24 *Graph comparing the price in U.S. dollars with a customer loyalty score (with rug representation)*

This rug indicates that the majority of customers in this example are in a tight band of loyalty scores, between about 1 and 3 on the x-axis, all of which offered the same set of prices, which are high (between 0.9 and 1.0 on the y-axis). The y-axis in this example may represent a pricing score, or the raw value of a customer in millions of dollars. The important aspect is to recognize that the pricing is high and is offered consistently to most of the customers in this example.

Based on what was shown in Figure 12-25, the team may decide to develop a new pricing model. Rather than offering static prices to customers regardless of their level of loyalty, a new pricing model might offer more dynamic price points to customers. In this visualization, the data shows the price increases as more of a curvilinear slope relative to the customer loyalty score. The rug at the bottom of the graph indicates that most customers remain between 1 and 3 on the x-axis, but now rather than offering all these customers the same price, the proposal suggests offering progressively higher prices as customer loyalty increases. In one sense, this may seem counterintuitive. It could be argued that the best prices should be offered to the most loyal customers. However, in reality, the opposite is often the case, with the most attractive prices being offered to the least loyal customers. The rationale is that loyal customers are less price sensitive and may enjoy the product and stay with it regardless of small fluctuations in price. Conversely, customers who are not very loyal may defect unless they are offered more attractive prices to stay. In other words, less loyal customers are more price sensitive. To address this issue, a new pricing model that accounts for this may enable an organization to maximize revenue and minimize attrition by offering higher prices to more loyal customers and lower prices to less loyal customers. Creating an iterative depicting the data visually allows the viewer to see these changes in a more concrete way than by looking at tables of numbers or raw values.

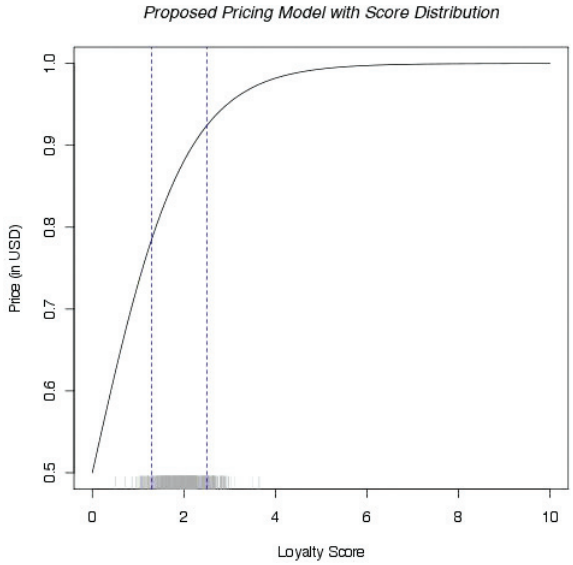


FIGURE 12-25 *New proposed pricing model compared to prices in U.S. dollars with rug*

Data scientists typically iterate and view data in many different ways, framing hypotheses, testing them, and exploring the implications of a given model. This case explores visual examples of pricing distributions, fluctuations in pricing, and the differences in price tiers before and after implementing a new model to optimize price. The visualization work illustrates how the data may look as the result of the model, and helps a data scientist understand the relationships within the data at a glance.

The resulting graph in the pricing scenario appears to be technical regarding the distribution of prices throughout a customer base and would be suitable for a technical audience composed of other data scientists. Figure 12-26 shows an example of how one may present this graphic to an audience of other data scientists or data analysts. This demonstrates a curvilinear relationship between price tiers and customer loyalty when expressed as an index. Note that the comments to the right of the graph relate to the precision of the price targeting, the amount of variability in robustness of the model, and the expectations of model speed when run in a production environment.

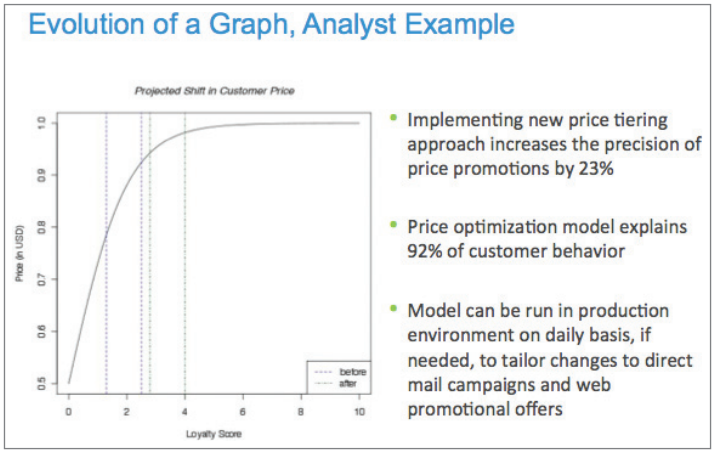


FIGURE 12-26 *Evolution of a graph, analyst example with supporting points*

Figure 12-27 portrays another example of the output from the price optimization project scenario, showing how one may present this to an audience of project sponsors. This demonstrates a simple bar chart depicting the average price per customer or user segment. Figure 12-27 shows a much simpler-looking visual than Figure 12-26. It clearly portrays that customers with lower loyalty scores tend to get lower prices due to targeting from price promotions. Note that the right side of the image focuses on the business impact and cost savings rather than the detailed characteristics of the model.

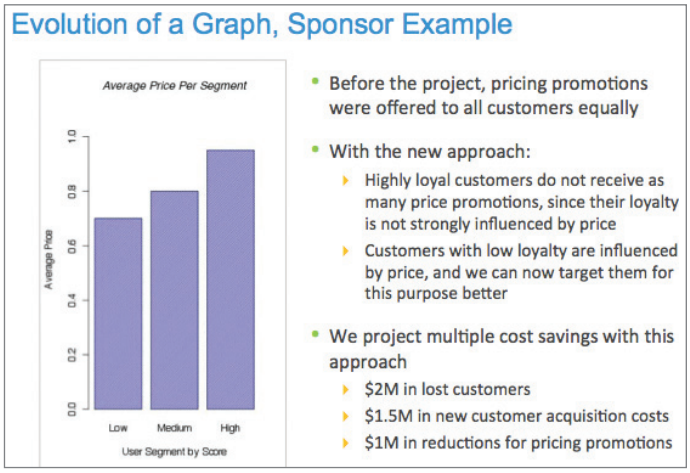


FIGURE 12-27 *Evolution of a graph, sponsor example*

The comments to the right side of the graphic in Figure 12-27 explain the impact of the model at a high level and the cost savings of implementing this approach to price optimization.

12.3.3 Common Representation Methods

Although there are many types of data visualizations, several fundamental types of charts portray data and information. It is important to know when to use a particular type of chart or graph to express a given kind of data. Table 12-3 shows some basic chart types to guide the reader in understanding that different types of charts are more suited to a situation depending on specific kinds of data and the message the team is attempting to portray. Using a type of chart for data it is not designed for may look interesting or unusual, but it generally confuses the viewer. The objective for the author is to find the best chart for expressing the data clearly so the visual does not impede the message, but rather supports the reader in taking away the intended message.

TABLE 12-3 *Common Representation Methods for Data and Charts*

Data for Visualization	Type of Chart
Components (parts of whole)	Pie chart
Item	Bar chart
Time series	Line chart
Frequency	Line chart or histogram
Correlation	Scatterplot, side-by-side bar charts

Table 12-3 shows the most fundamental and common data representations, which can be combined, embellished, and made more sophisticated depending on the situation and the audience. It is recommended

that the team consider the message it is trying to communicate and then select the appropriate type of visual to support the point. Misusing charts tends to confuse an audience, so it is important to take into account the data type and desired message when choosing a chart.

Pie charts are designed to show the components, or parts relative to a whole set of things. A pie chart is also the most commonly misused kind of chart. If the situation calls for using a pie chart, employ it only when showing only 2–3 items in a chart, and only for sponsor audiences.

Bar charts and line charts are used much more often and are useful for showing comparisons and trends over time. Even though people use vertical bar charts more often, horizontal bar charts allow an author more room to fit the text labels. Vertical bar charts tend to work well when the labels are small, such as when showing comparisons over time using years.

For frequency, histograms are useful for demonstrating the distribution of data to an analyst audience or to data scientists. As shown in the pricing example earlier in this chapter, data distributions are typically one of the first steps when visualizing data to prepare for model planning. To qualitatively evaluate correlations, scatterplots can be useful to compare relationships among variables.

As with any presentation, consider the audience and level of sophistication when selecting the chart to convey the intended message. These charts are simple examples but can easily become more complex when adding data variables, combining charts, or adding animation where appropriate.

12.3.4 How to Clean Up a Graphic

Many times software packages generate a graphic for a dataset, but the software adds too many things to the graphic. These added visual distractions can make the visual appear busy or otherwise obscure the main points that are to be made with the graphic. In general, it is a best practice to strive for simplicity when creating graphics and data visualization graphs. Knowing how to simplify graphics or clean up a messy chart is helpful for conveying the key message as clearly as possible. Figure 12-28 portrays a line chart with several design problems.

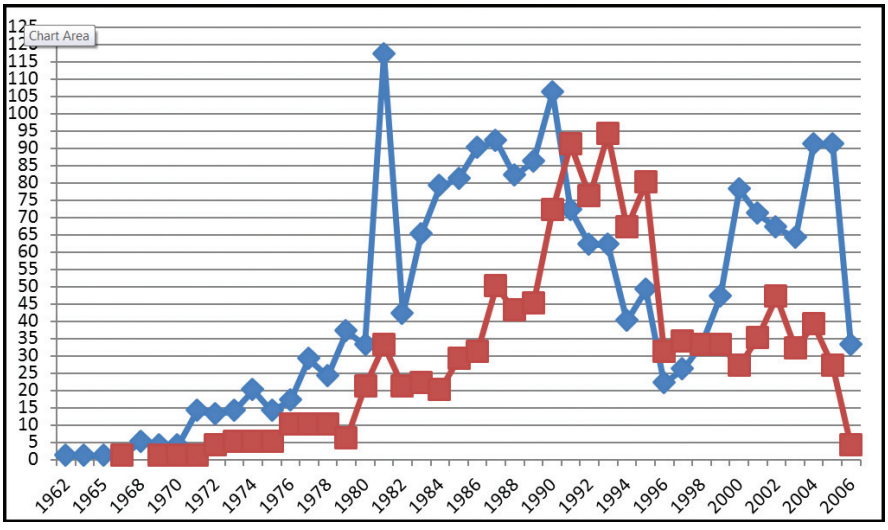


FIGURE 12-28 *How to clean up a graphic, example 1 (before)*

How to Clean Up a Graphic

The line chart shown in Figure 12-28 compares two trends over time. The chart looks busy and contains a lot of chart junk that distracts the viewer from the main message. *Chart junk* refers to elements of data visualization that provide additional materials but do not contribute to the data portion of the graphic. If chart junk were removed, the meaning and understanding of the graphic would not be diminished; it would instead be made clearer. There are five main kinds of "chart junk" in Figure 12-28:

- **Horizontal grid lines:** These serve no purpose in this graphic. They do not provide additional information for the chart.
- **Chunky data points:** These data points represented as large square blocks draw the viewer's attention to them but do not represent any specific meaning aside from the data points themselves.
- **Overuse of emphasis colors in the lines and border:** The border of the graphic is a thick, bold line. This forces the viewer's attention to the perimeter of the graphic, which contains no information value. In addition, the lines showing the trends are relatively thick.
- **No context or labels:** The chart contains no legend to provide context as to what is being shown. The lines also lack labels to explain what they represent.
- **Crowded axis labels:** There are too many axis labels, so they appear crowded. There is no need for labels on the y-axis to appear every five units or for values on the x-axis to appear every two units. Shown in this way, the axis labels distract the viewer from the actual data that is represented by the trend lines in the chart.

The five forms of chart junk in Figure 12-28 are easily corrected, as shown in Figure 12-29. Note that there is no clear message associated with the chart and no legend to provide context for what is shown in Figure 12-28.

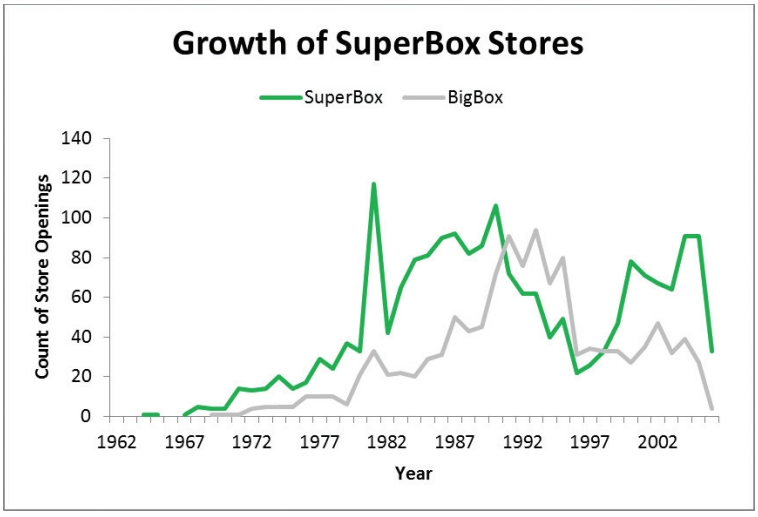


FIGURE 12-29 *How to clean up a graphic, example 1 (after)*

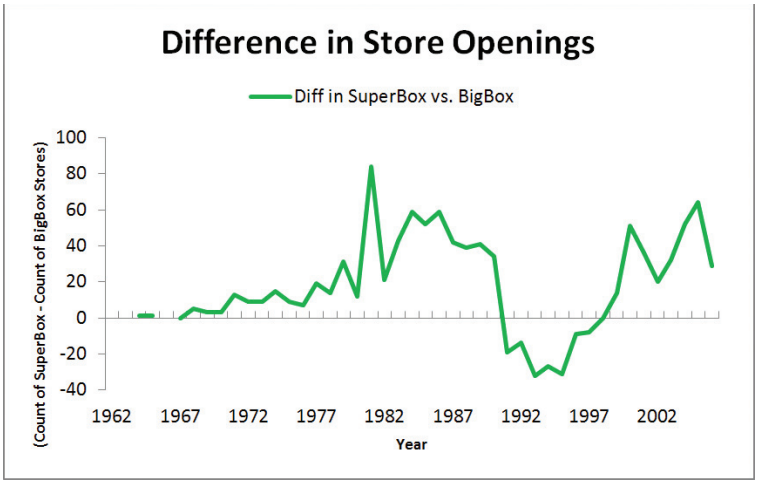


FIGURE 12-30 *How to clean up a graphic, example 1 (alternate "after" view)*

Figures 12-29 and 12-30 portray two examples of cleaned-up versions of the chart shown in Figure 12-28. Note that the problems with chart junk have been addressed. There is a clear label and title for each chart to reinforce the message, and color has been used in ways to highlight the point the author is trying to make. In Figure 12-29, a strong, green color is shown to represent the count of SuperBox stores, because this is where the viewer's focus should be drawn, whereas the count of BigBox stores is shown in a light gray color.

In addition, note the amount of white space being used in each of the two charts shown in Figures 12-29 and 12-30. Removing grid lines, excessive axes, and the visual noise within the chart allows clear contrast between the emphasis colors (the green line charts) and the standard colors (the lighter gray of the BigBox stores). When creating charts, it is best to draw most of the main visuals in standard colors, light tones, or color shades so that stronger emphasis colors can highlight the main points. In this case, the trend of BigBox stores in light gray fades into the background but does not disappear, while making the SuperBox stores trend in a darker gray (bright green in the online chart) makes it prominent to support the message the author is making about the growth of the SuperBox stores.

An alternative to Figure 12-29 is shown in Figure 12-30. If the main message is to show the difference in the growth of new stores, Figure 12-30 can be created to further simplify Figure 12-28 and graph only the difference between SuperBox stores compared to regular BigBox stores. Two examples are shown to illustrate different ways to convey the message, depending on what it is the author of these charts would like to emphasize.

How to Clean Up a Graphic, Second Example

Another example of cleaning up a chart is portrayed in Figure 12-31. This vertical bar chart suffers from more of the typical problems related to chart junk, including misuse of color schemes and lack of context.

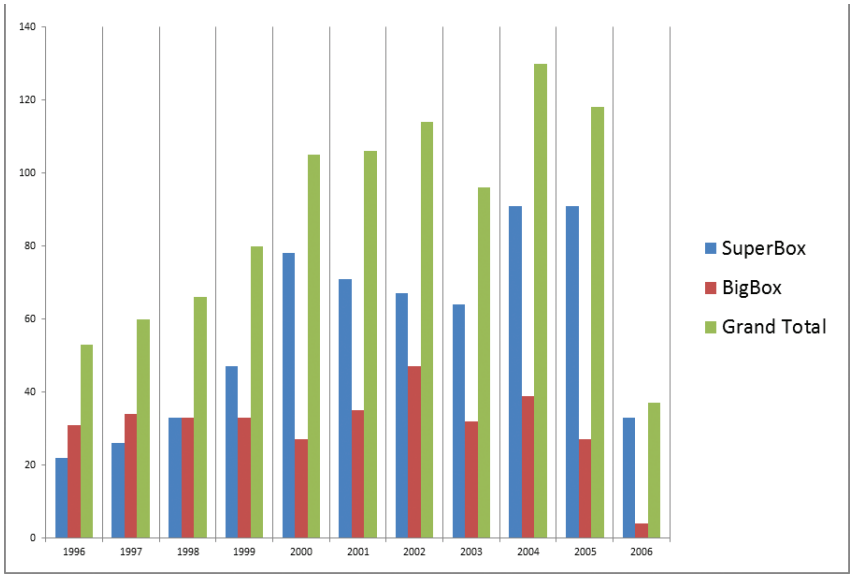


FIGURE 12-31 *How to clean up a graphic, example 2 (before)*

There are five main kinds of chart junk in Figure 12-31:

- **Vertical grid lines:** These vertical grid lines are not needed in this graphic. They provide no additional information to help the viewer understand the message in the data. Instead, these vertical grid lines only distract the viewer from looking at the data.
- **Too much emphasis color:** This bar chart uses strong colors and too much high-contrast dark gray-scale. In general, it is best to use subtle tones, with a low contrast gray as neutral color, and then emphasize the data underscoring the key message in a dark tone or strong color.
- **No chart title:** Because the graphic lacks a chart title, the viewer is not oriented to what he is viewing and does not have proper context.
- **Legend at right restricting chart space:** Although there is a legend for the chart, it is shown on the right side, which causes the vertical bar chart to be compressed horizontally. The legend would make more sense placed across the top, above the chart, where it would not interfere with the data being expressed.
- **Small labels:** The horizontal and vertical axis labels have appropriate spacing, but the font size is too small to be easily read. These should be slightly larger to be easily read, while not appearing too prominent.

Figures 12-32 and 12-33 portray two examples of cleaned-up versions of the chart shown in Figure 12-31. The problems with chart junk have been addressed. There is a clear label and title for each chart to reinforce the message, and appropriate colors have been used in ways to highlight the point the author is trying to make. Figures 12-32 and 12-33 show two options for modifying the graphic, depending on the main point the presenter is trying to make.

Figure 12-32 shows strong emphasis color (dark blue) representing the SuperBox stores to support the chart title: Growth of SuperBox Stores.

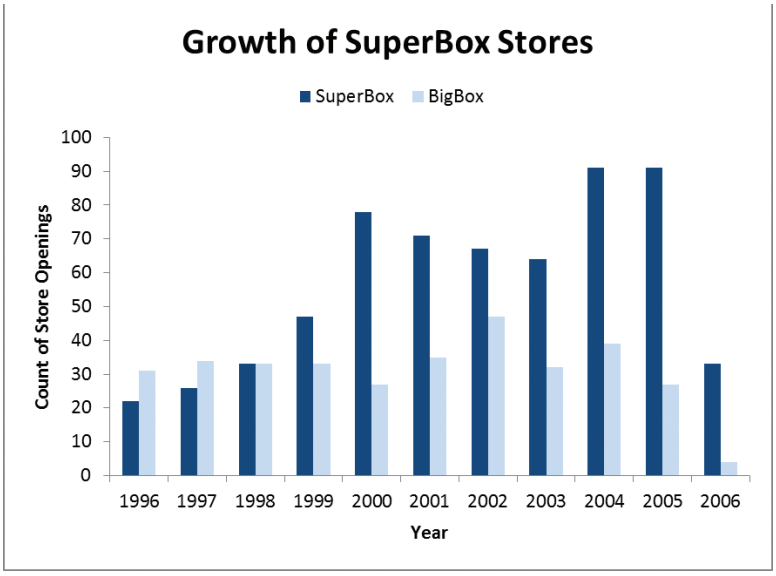


FIGURE 12-32 *How to clean up a graphic, example 2 (after)*

Suppose the presenter wanted to talk about the total growth of BigBox stores instead. A line chart showing the trends over time would be a better choice, as shown in Figure 12-33.

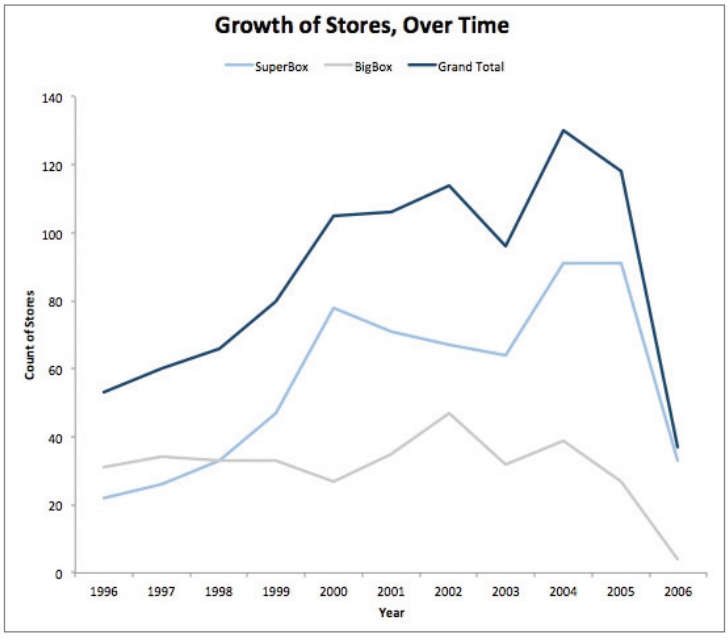


FIGURE 12-33 *How to clean up a graphic, example 2 (alternate view of "after")*

In both cases, the noise and distractions within the chart have been removed. As a result, the data in the bar chart for providing context has been deemphasized, while other data has been made more prominent because it reinforces the key point as stated in the chart's title.

12.3.5 Additional Considerations

As stated in the previous examples, the emphasis should be on simplicity when creating charts and graphs. Create graphics that are free of chart junk and utilize the simplest method for portraying graphics clearly. The goal of data visualization should be to support the key messages being made as clearly as possible and with few distractions.

Similar to the idea of removing chart junk is being cognizant of the data-ink ratio. *Data-ink* refers to the actual portion of a graphic that portrays the data, while *non-data ink* refers to labels, edges, colors, and other decoration. If one imagined the ink required to print a data visualization on paper, the data-ink ratio could be thought of as (data-ink)/(total ink used to print the graphic). In other words, the greater the ratio of data-ink in the visual, the more data rich it is and the fewer distractions it has [4].

Avoid Using Three-Dimensions in Most Graphics

One more example where people typically err is in adding unnecessary shading, depth, or dimensions to graphics. Figure 12-34 shows a vertical bar chart with two visible dimensions. This example is simple and easy to understand, and the focus is on the data, not the graphics. The author of the chart has chosen to highlight the SuperBox stores in a dark blue color, while the BigBox bars in the chart are in a lighter blue. The title is about the growth of SuperBox stores, and the SuperBox bars in the chart are in a dark, high-contrast shade that draws the viewer's attention to them.

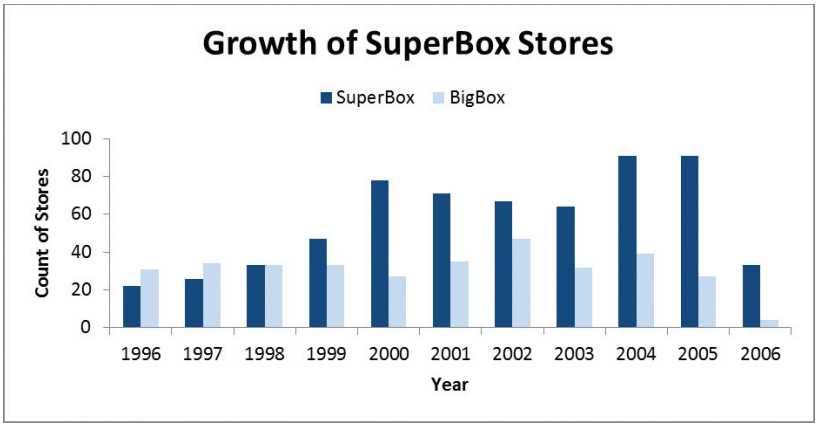


FIGURE 12-34 *Simple bar chart, with two dimensions*

Compare Figure 12-34 to Figure 12-35, which shows a three-dimensional chart. Figure 12-35 shows the original bar chart at an angle, with some attempt at showing depth. This kind of three-dimensional perspective makes it more difficult for the viewer to gauge the actual data and the scaling becomes deceptive.

Three-dimensional charts often distort scales and axes, and impede viewer cognition. Adding a third dimension for depth in Figure 12-35, does not make it fancier, just more difficult to understand.

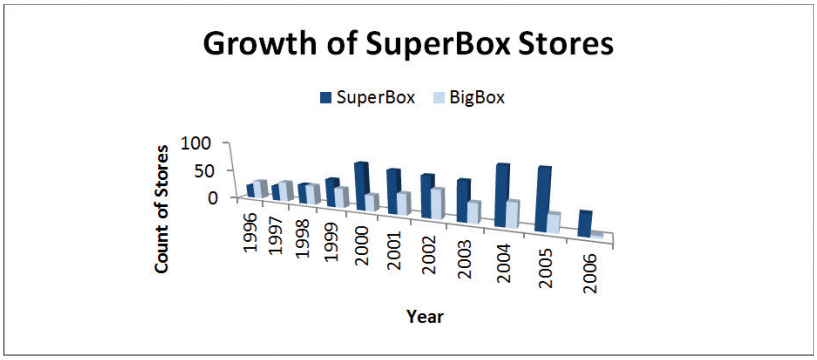


FIGURE 12-35 *Misleading bar chart, with three dimensions*

The charts in Figures 12-34 and 12-35 portray the same data, but it is more difficult to judge the actual height of the bars in Figure 12-35. Moreover, the shadowing and shape of the chart cause most viewers to spend time looking at the perspective of the chart rather than the height of the bars, which is the key message and purpose of this data visualization.

Summary

Communicating the value of analytical projects is critical for sustaining the momentum of a project and building support within organizations. This support is instrumental in turning a successful project into a system or integrating it properly into an existing production environment. Because an analytics project may need to be communicated to audiences with mixed backgrounds, this chapter recommends creating four deliverables to satisfy most of the needs of various stakeholders.

- A presentation for a project sponsor
- A presentation for an analytical audience
- Technical specification documents
- Well-annotated production code

Creating these deliverables enables the analytics project team to communicate and evangelize the work that it did, whereas the code and technical documentation assists the team that wants to implement the models within the production environment.

This chapter illustrates the importance of selecting clear and simple visual representations to support the key points in the final presentations or for portraying data. Most data representations and graphs can be improved by simply removing the visual distractions. This means minimizing or removing chart junk, which distracts the viewer from the main purpose of a chart or graph and does not add information value.

Following several common-sense principles about minimizing distractions in slides and visualizations, communicating clearly and simply, using color in a deliberate way, and taking time to provide context addresses most of the common problems in charts and slides. These few guidelines support the creation of crisp, clear visuals that convey the key messages.

In most cases, the best data visualizations use the simplest, clearest visual to illustrate the key point. Avoid unnecessary embellishment and focus on trying to find the best, simplest method for transmitting the message. Context is critical to orient the viewer to a chart or graph, because people have immediate reactions to imagery on a precognitive level. To this end, make sure to employ thoughtful use of color and orient the viewer with scales, legends, and axes.

Exercises

1. Describe four common deliverables for an analytics project.
2. What is the focus of a presentation for a project sponsor?
3. Give examples of appropriate charts to create in a presentation for other data analysts and data scientists as part of a final presentation. Explain why the charts are appropriate to show each audience.
4. Explain what types of graphs would be appropriate to show data changing over time and why.
5. As part of operationalizing an analytics project, which deliverable would you expect to provide to a Business Intelligence analyst?

References and Further Reading

Following are additional references to learn more about best practices for giving presentations.

- ***Say It with Charts*, by Gene Zelazny [3]:** Simple reference book on how to select the right graphical approach for portraying data and for ensuring the message is clearly conveyed in presentations.
- ***Pyramid Principle*, by Barbara Minto [5]:** Minto pioneered the approach for constructing logical structures for presentations in threes: three sections to the presentations, each with three main points. This teaches people how to weave a story out of the disparate pieces.
- ***Presentation Zen*, by Garr Reynolds [6]:** Teaches how to convey ideas simply and clearly and use imagery in presentations. Shows many before and after versions of graphics and slides.
- ***Now You See It*, by Stephen Few [4]:** Provides many examples for matching the appropriate kind of data visualization to a given dataset.

Bibliography

[1] N. Yau, "flowingdata.com" [Online]. Available: `http://flowingdata.com`.
[2] N. Yau, *Visualize This*, Indianapolis: Wiley, 2011.
[3] G. Zelazny, *Say It with Charts: The Executive's Guide to Visual Communication*, McGraw-Hill, 2001.
[4] S. Few, *Now You See It: Simple Visualization Techniques for Quantitative Analysis*, Analytics Press, 2009.
[5] B. Minto, *The Minto Pyramid Principle: Logic in Writing, Thinking, and Problem Solving*, Prentice Hall, 2010.
[6] G. Reynolds, *Presentation Zen: Simple Ideas on Presentation Design and Delivery*, Berkeley: New Riders, 2011.

Index

Numbers & Symbols

A

B

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

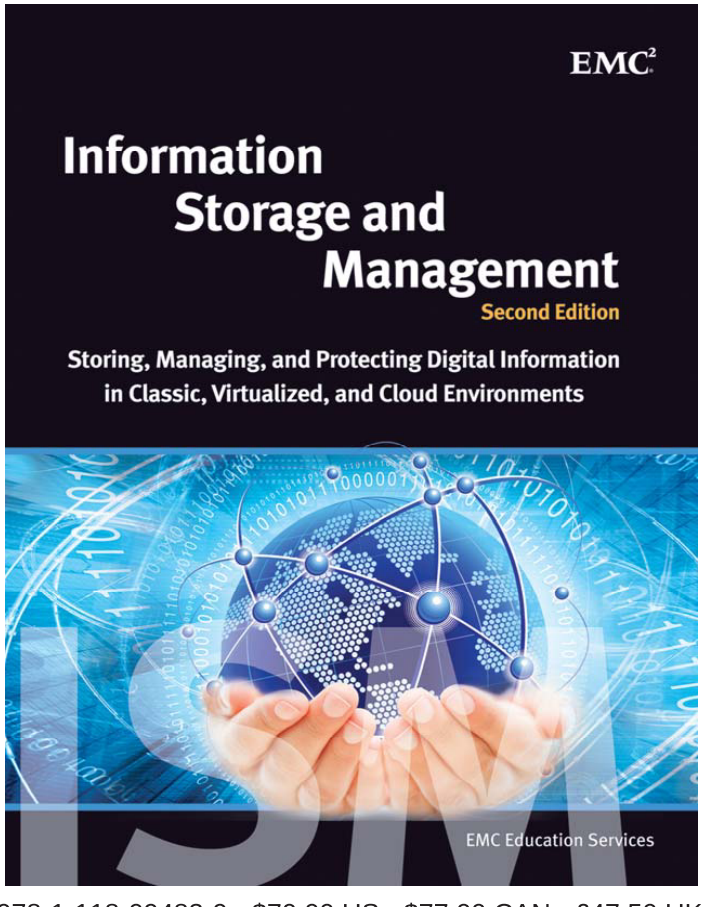
V

W

X-Z

A comprehensive book on information storage and management by the EMC's Education Services, a global technology leader.

More than ever, the IT industry is challenged with employing and developing highly skilled technical professionals with storage technology expertise across classic, virtualized, and cloud environments. This book covers concepts, principles, and deployment considerations across technologies that are used for storing and managing information.



Key Technology Strategies for Classic, Virtualized, and Cloud Environments:

- Challenges and Solutions for Data Storage and Management
- Intelligent Storage, Object-Based Storage, and Unified Storage
- Storage Networking, Federation, and Protocols
- Backup, Recovery, Deduplication, and Archive
- Business Continuity and Disaster Recovery
- Cloud Computing and Converged Infrastructure
- Storage Security and Managing Storage Infrastructure

978-1-118-09483-9 · $70.00 US · $77.00 CAN · £47.50 UK

WILEY

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.