Paper 235-2010

Effective Visualization Techniques for Data Discovery and Analysis

Chuck Pirrello, SAS Institute, Cary, NC

ABSTRACT

One of the best ways to understand data is through data visualization. The art of presenting data visually has matured from static graphs and dashboard widgets to interactive, animated graphs. Adding advanced analytics to support these graphs provides a competitive edge to companies by helping them better explore and understand their data, predict potential outcomes and decide with confidence.

This session will discuss the power of visualization for analyzing data and spotting trends you can act on. Effective visual techniques will be demonstrated showing how users can better explore and understand their data, discover trends and patterns, and communicate their findings to a non-technical, non-analytical audience.

INTRODUCTION

Most information researchers agree that the volume of electronically stored data is doubling every 18 months. Some say it's happening even faster. Regardless of the actual time frame, data is being generated faster than it can be consumed and digested. Companies and individuals are constantly being challenged to make decisions more quickly with less time to analyze data that is required to support their conclusions and proposed courses of action. This presentation offers best practices for analyzing and displaying data graphically, using techniques that enhance clarity and shorten the path to discovery.

VISUAL ANALYTICS

The components of visual analytics are interactive graphics and advanced analytics. Those elements combine to create statistical models that allow analysts to synthesize large amounts of data, detect trends and patterns, and help discover important and unexpected content. The resulting insights permit analysts to make high-quality human judgments more quickly.

"Visual analytics is the science of analytical reasoning facilitated by interactive visual interfaces … [and] seeks to marry techniques from information visualization with techniques from computational transformation and analysis of data. The computer finds patterns in the information and organizes it in ways that are meant to be revealing to the analyst. The analyst supplies his or her knowledge in ways that help the computer refine and organize information more appropriately. Working together, they are much more powerful than each one working separately." (*Illuminating the Path: The Research and Development Agenda for Visual Analytics*, National Visualization and Analytics Center, 2004.)

STATIC GRAPHS

Static graphs are appropriate when a snapshot or series of snapshots can convey the intended information. Care should be taken not to overload the graph with too much decoration that can conceal the data. Conversely, not showing enough information in a graph also can dilute its effectiveness.

The failure of poorly prepared graphs was painfully and tragically evident with the Challenger O-ring incident in 1986, which resulted in the explosion of Space Shuttle Challenger shortly after liftoff.

"Thirteen charts were prepared to make the case [about the danger posed by the O-rings], but the charts were unconvincing. The engineers correctly identified the O-ring failure at low temperatures, but the displays chosen to present the evidence did not adequately show the cause of the failure and overcome the bias of decision makers. Displays obscured the data, and the wrong decision to launch was made. The consequences were tragic."

(*Illuminating the Path: The Research and Development Agenda for Visual Analytics*, National Visualization and Analytics Center, 2004.)

GUIDELINES FOR CREATING STATIC GRAPHS:

It's all about the data

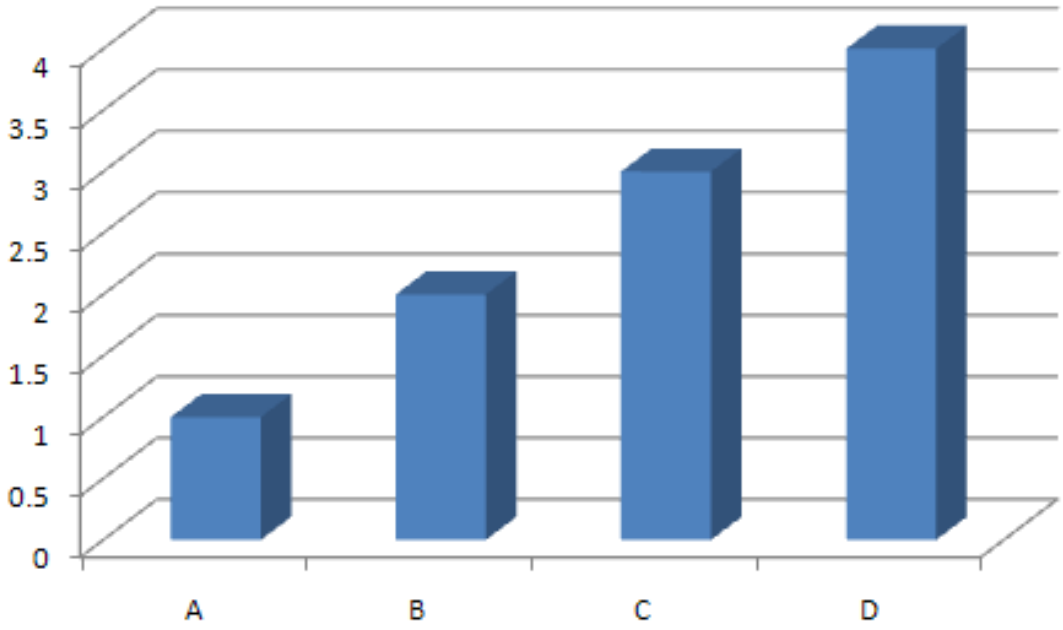
All too often, users get hung up on the "bling" factor in creating graphs. They think that if they wow the viewer with elaborate, attention-getting images, it will help them see the results more clearly. In fact, overblown images often obscure the meaning of the data. Graphs should bring the data to life, not conceal it.

2-D or 3?

Most people use a third dimension to dress up a graph rather than to show another variable. Such things as 3-D bars with shadows do nothing more than add a visual effect that, while possibly more pleasing to the eye, contributes no new information. Avoid using 3-D graphs merely to imitate depth or shadows. As a rule, try first to create the visual in 2-D. If all the information can be clearly conveyed, stay with that design.

Sometimes a third dimension is used to represent a third variable. However, you can often convey the same information on a 2-D chart by varying the color or size of the objects shown.

If you feel you must use 3-D, do so only with interactive graphs, not static ones. The viewer should be able to rotate the graph to see all the information, even though you risk concealing some information by revealing other information.



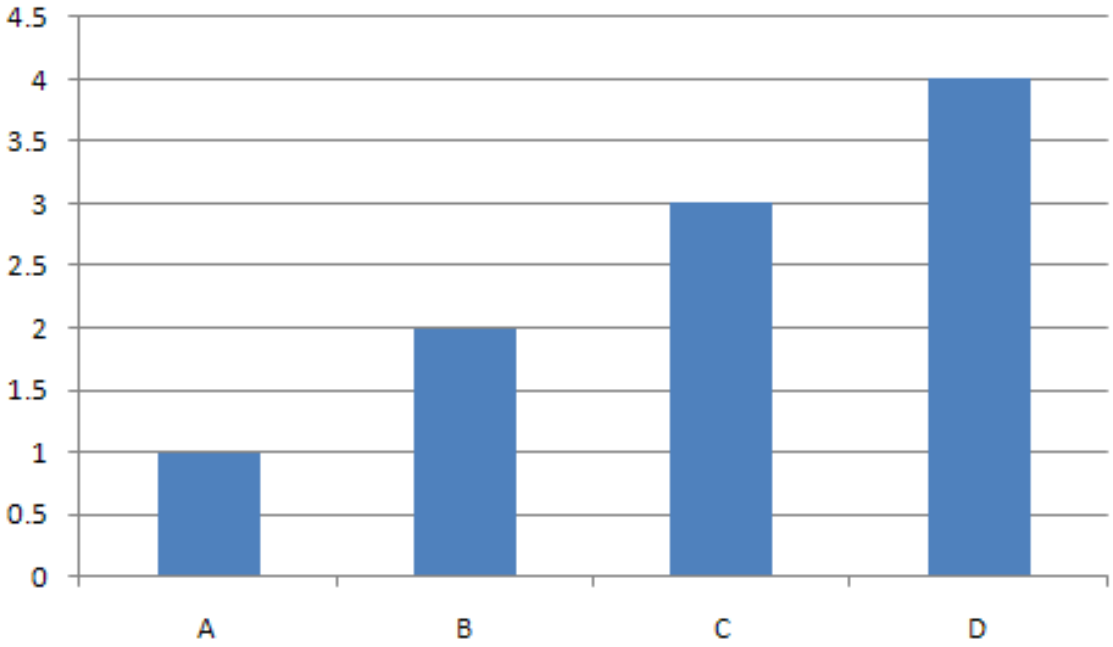


Figure 1. Adding a third dimension frequently adds little value for the viewer. In the top image here, the 3-D effect is misleading. The values are the same for both graphs. The second graph is merely displayed in two dimensions.

Plan your color scheme

Beauty may be in the eye of the beholder, but there are some tried and true "eye-pleasing" colors that appeal to most people and don't act as a distraction. For guidelines on color schemes, go to http://colorbrewer2.org/.

Conserve ink

Too much ink can conceal the actual data. Use light gray lines for axes and tick marks, and don't show unnecessary tick marks on graphs. As a basic guideline, use muted colors for all graphic elements other than the data and its graphical representation (i.e., bars, lines, etc.).

Use the right type of graph

When comparing the results among several entities, bar graphs typically provide the best visual. Bar graphs convey differences mainly through the differences in length, color or size of the bars. Care should be taken not to overload bar graphs with too many colors or varied bar widths for these attributes make it more difficult to see the relative length of the bars.

Line graphs are the best choice for data that is typically viewed over time, because they allow viewers to see how values have progressed and where each one may be headed.

How many is too many?

One common mistake is trying to put all the data into one graph, resulting in a graph that is too cluttered to make any sense. However, there is also a danger in creating too many graphs that must be viewed one at a time. This approach risks taxing the viewer's memory, making it difficult for him or her to grasp important details and comparisons. One way to solve this dilemma is by creating a display that includes several small graphs.

"Small multiples," or trellis displays, use a series of small graphs along the same x and y scales to present information in such a way that the viewer can read the data from graph to graph in a continuous fashion. Such a display allows the viewer to consume more information and facilitates comparisons.

Small multiples should exhibit the following characteristics:

- Individual graphs differ only in terms of the data that they display. Each graph displays a subset of a single larger data set.
- Graphs are identical in shape, type and size, and they share the same categorical and quantitative scales.
- Graphs can be arranged horizontally, vertically or in a matrix.
- Graphs are sequenced in a meaningful order. (*Now You See It: Simple Visualization Techniques for Quantitative Analysis*, Stephen Few, Analytics Press, 2009.)

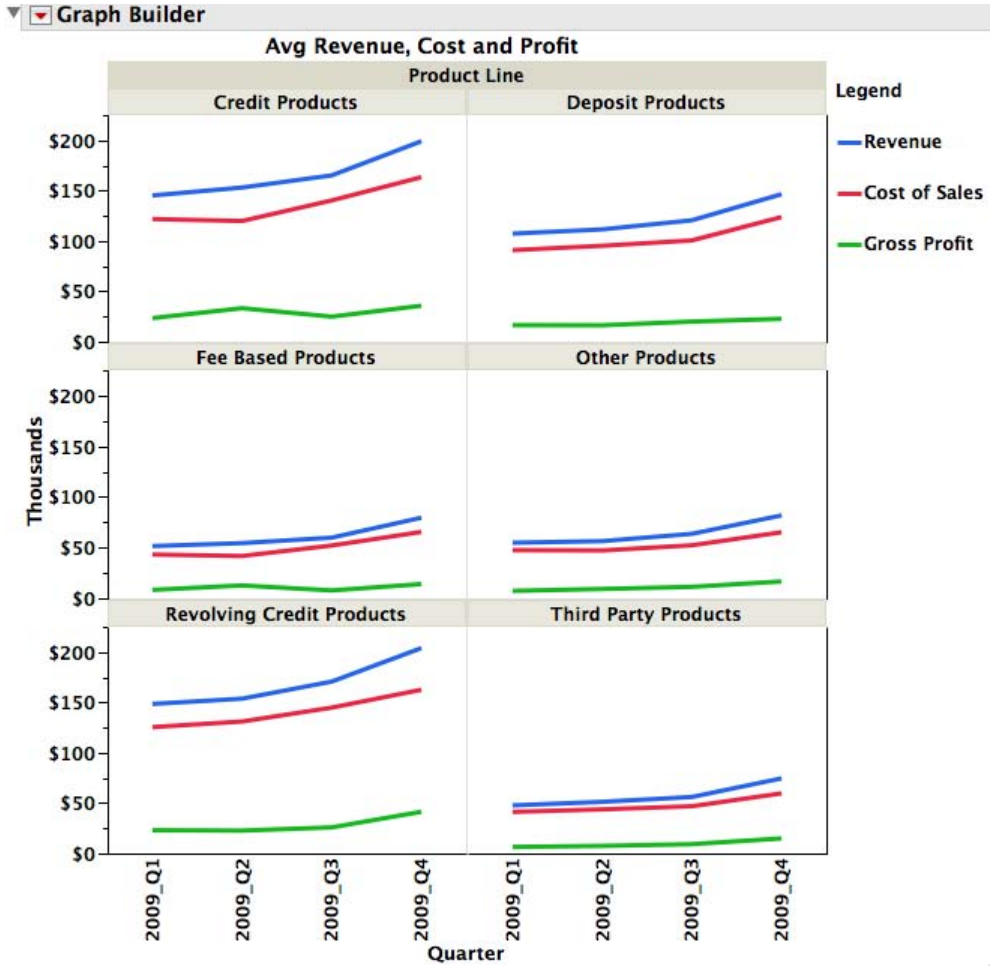


Figure 2. A small multiples graph facilitates comparisons.

Keep time axes honest

A frequent misuse of time axes is to skew the tick marks to exaggerate or downplay the significance of the data.

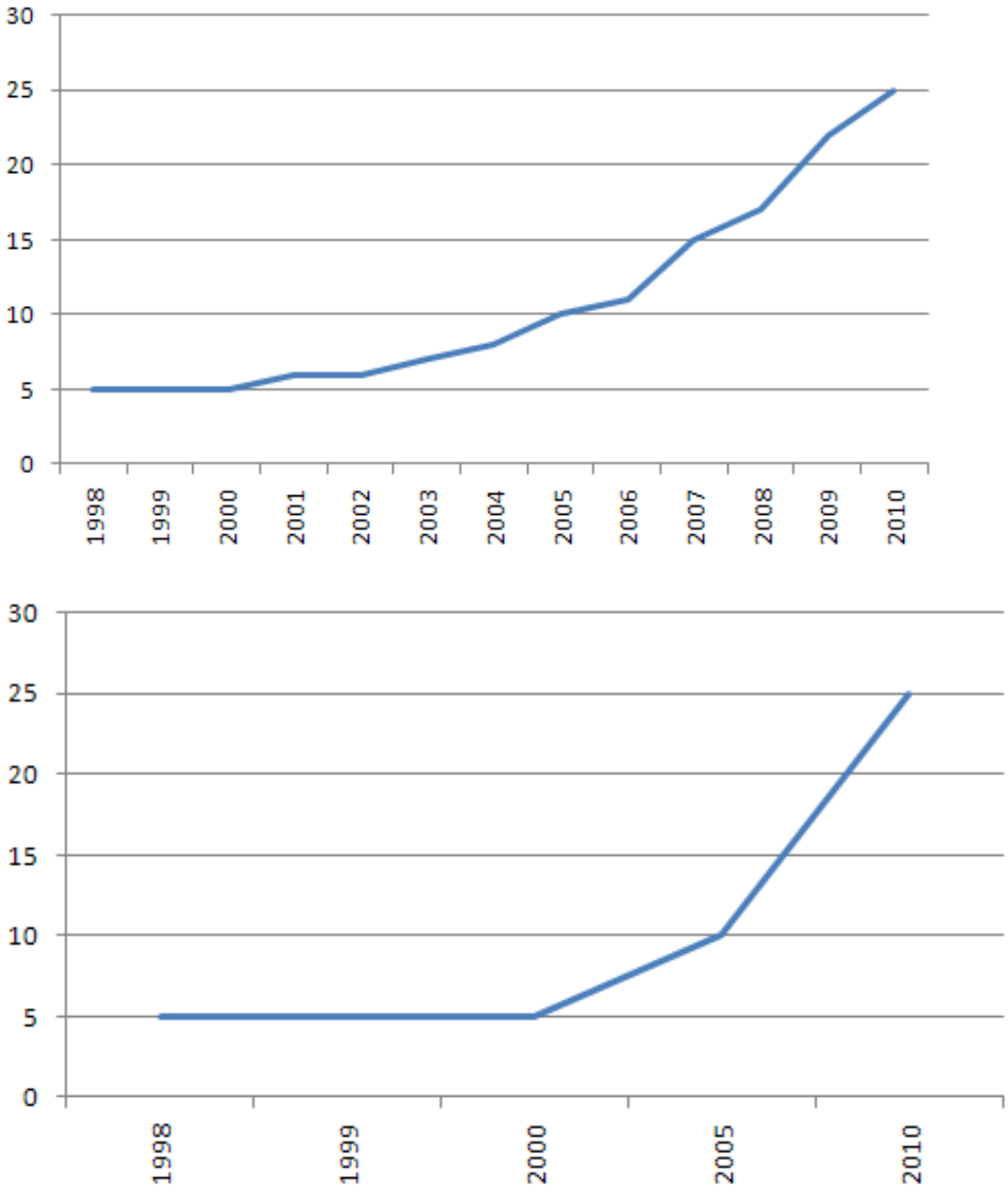


Figure 3. The top graph shows a more gradual increase over time. The second graph misleads the viewer into thinking the rise over time is much more dramatic by changing midstream from yearly increments to five-year increments.

INTERACTIVE GRAPHS

While static graphs are commonly used to convey findings that result from analysis, they do not usually yield new insights. Interactive graphs are much more likely than static graphs to provide insights that lead to discovery.

Static graphs can quickly become cluttered when large amounts of information are required for understanding and decision-making. Interactive graphs let analysts “spoon feed” the audience, revealing details progressively instead of all at once so that the information becomes more “digestible.”

Visual representations should invite the user and audience to explore the data, but exploration is not possible if they cannot interact with the data. The ability to isolate and reorganize information appropriately leads to better understanding of trends and anomalies, and engages them in the analytic reasoning process. It is through these interactions that the analyst achieves insight.

Interaction takes four basic forms: selecting desired data, reorganizing data, zooming and data filtering.

1. In an interactive graph, selecting an object in one graph results in a selection in a linked graph, as seen here.

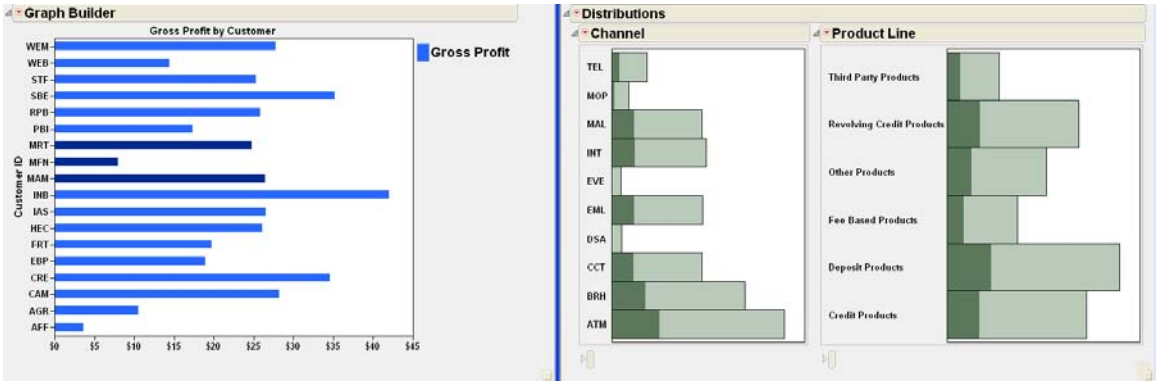


Figure 4. Selecting bars in the bar graph on the left causes the distribution graph on the right to reflect the selection.

2. Reorganizing data means letting the user redefine or reorient the data. This type of interaction can include the x and y axes, the legend, and any other dimensions the graph is capable of displaying.
3. Zooming allows the user to enlarge any portion of a graph to see more detail.
4. Filtering allows the user to narrow the view of the data, as in the following graph.

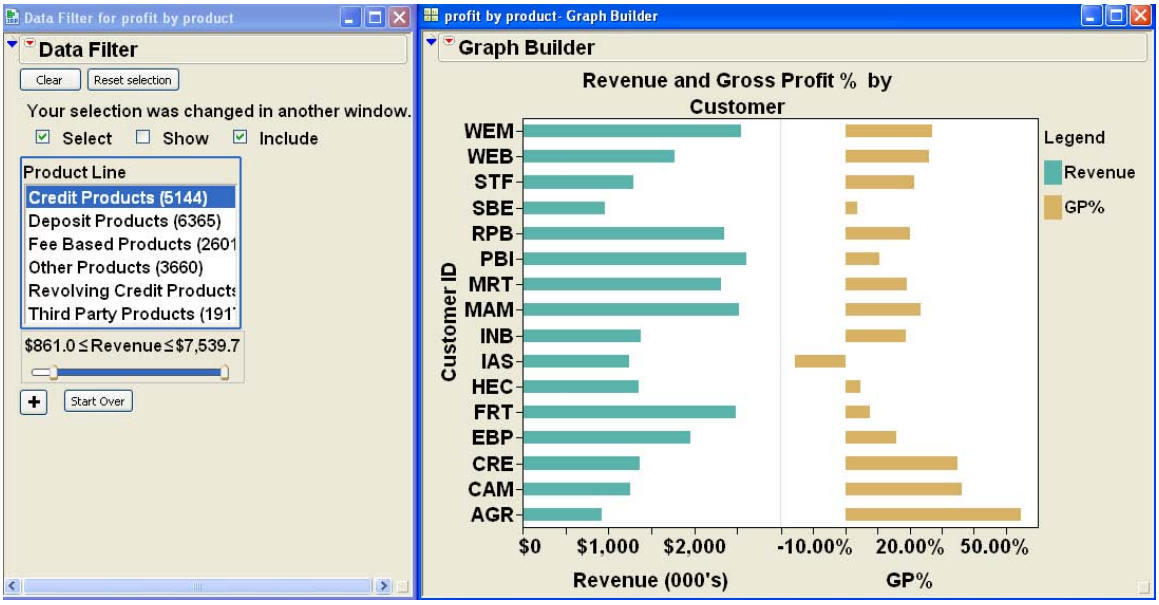


Figure 5. The filter on the left governs what data is displayed in the graph on the right.

Interactive graphs should be created with the same visual guidelines as static graphs. Their interactions and results should be displayed immediately to provide the analyst with the ability to navigate through data seamlessly, making it easier to spot trends and patterns. The ability to see data unfold before one's eyes can provide insights that static data simply cannot convey.

SUPPORTING ANALYTICS

Visuals that help users analyze their data transcend the capabilities of traditional graphs. They are navigational vehicles that let users peer into their data. Those that provide built-in methods for synthesizing and processing data offer analysts a deeper understanding of what the data represents and how future results might unfold.

For example, the ability to segment data based on similar results over time, or to predict potential outcomes by varying a number of inputs that contribute to a desired output (known as Monte Carlo simulation), enables analysts to move beyond merely dissecting data to gaining valuable insights into cause-and-effect relationships. A better understanding of root causes can facilitate decisions that are more likely to achieve desired results.

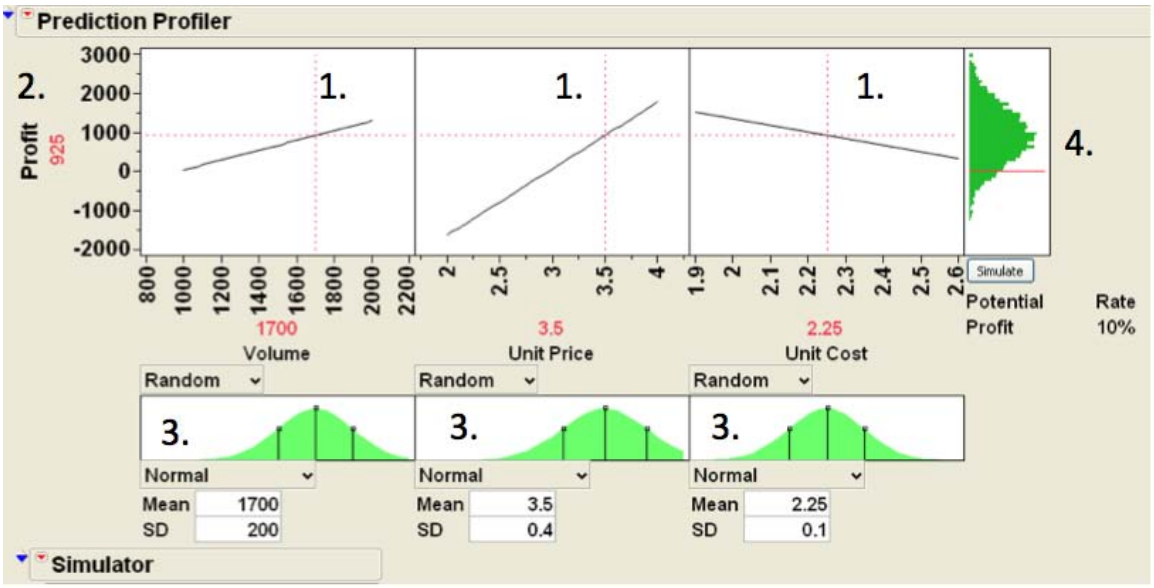


Figure 6. The Prediction Profiler in JMP® shows the effect of selected inputs on the desired output.

The display in Figure 6 uses Monte Carlo simulation to predict the potential outcomes of targeted values. The windows marked with a 1 represent each variable that contributes to the desired result: profit (2). A user can adjust the values of each input variable (1) by clicking and dragging the vertical red dotted line or by simply entering the value (shown in red below each variable display box.). As the user changes the input variables, the desired result is recalculated and displayed immediately. To use the Monte Carlo simulation, the user marks each variable as random or fixed, selects the distribution method (Normal is shown in this example), enters the number of iterations (5,000 or more) to be performed, and clicks on the SIMULATE button. The display window (4) will show the results of the simulation. In this example, values above the red line in display window 4 indicate profit, and values below the line indicate loss. The rate shows the percentage of the simulated outcomes that resulted in a loss (i.e., the potential percentage of transactions that will not generate a profit). The value of this interactive graph is the immediacy with which scenarios can be generated and displayed. Its speed in showing results makes it an invaluable tool in the decision-making process.

PRESENTATION

The ultimate goal of visual analytics is not creation of the graphics themselves, but the resulting insights that inform decisions based on the findings. The ultimate value of a graph is not how well it is designed, but rather how well it conveys a call to action.

There are three questions an analyst must ask when preparing to present findings:

1. Do I have a set of visuals that clearly show my findings? That is, can a viewer with no more knowledge than what is presented here understand what the visuals show?
2. Is the level of detail in my graph appropriate for my audience? Showing too much detail might overwhelm viewers, while showing too little could fail to convince them. One benefit of an interactive graph is that the analyst can start with the minimum amount of data and reveal more as needed by altering the graph, filtering and zooming.
3. Do I have a clear, concise narrative to accompany my visuals? This question is often overlooked. All presenters need to be good storytellers. It's the only way their work can come alive and convince the audience of the merits of their conclusion, persuading the audience to take action. The narrative must be thorough, but not too detailed. It must be appropriate for the audience, and it should always start with a statement of why the audience should listen. This statement should include enough information about the presenter to establish her credibility, if necessary. A valuable guide to storytelling is *The Story Factor: Inspiration, Influence, and Persuasion Through the Art of Storytelling*, by Annette Simmons.

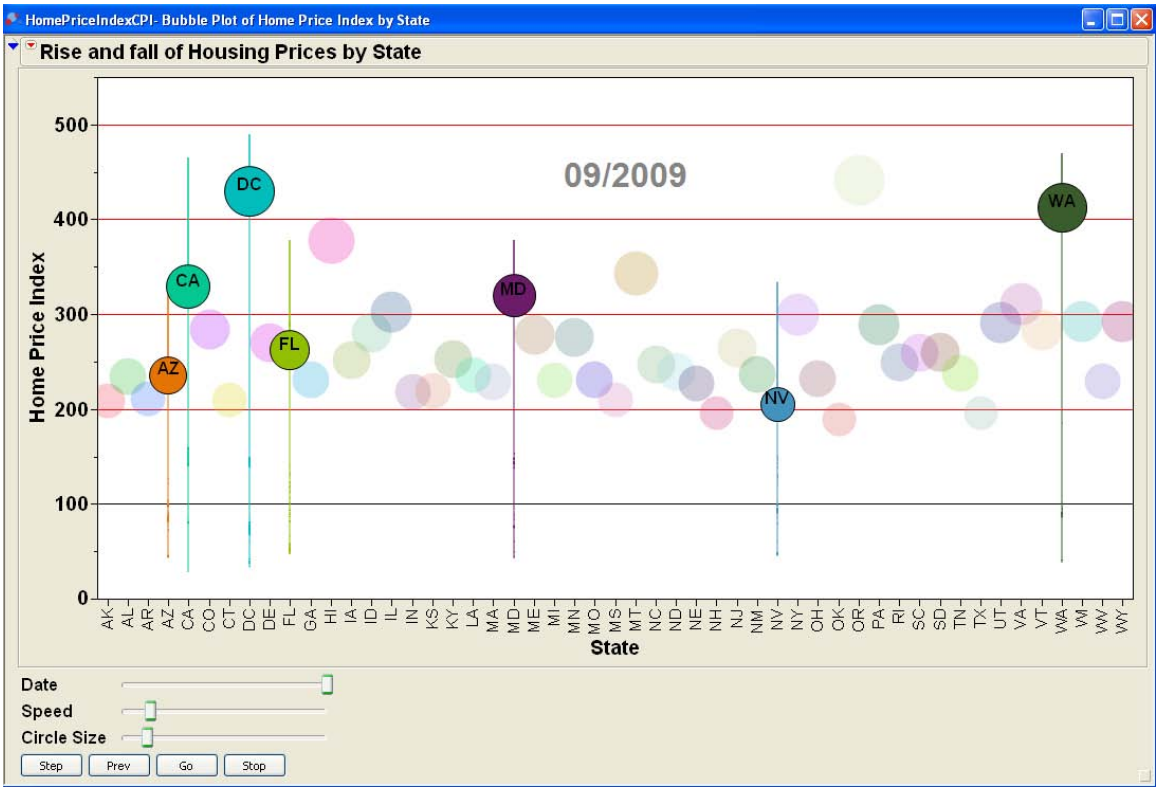


Figure 7. This bubble plot shows fluctuations in home prices over time. When animated, it shows how prices changed between 1975 and 2009.

Animated graphs require a good narration to be well understood. Once put in motion, the visuals need to reveal insights or guide viewers to areas of greatest interest. In the example shown in Figure 7, each bubble represents a

state and the rise and fall of housing prices there over time. When put in motion, each bubble rises over time, showing the pace and direction of price fluctuations.

CONCLUSION

Properly prepared graphs promote understanding of large amounts of data. Graphs that allow for user interaction and that are supported by underlying analytics offer even better understanding and insight. Care must be taken not to clutter graphs with unnecessary decoration, which can conceal the data that they are intended to surface. The analyst must match the data to be analyzed with the appropriate graph type. However, even well designed, interactive graphs can fail to achieve their goal if they are not accompanied by compelling narratives that not only explain, but also illuminate. In addition to good analytical skills, data analysts must also hone their design and storytelling skills to effectively create and deliver their findings.

RECOMMENDED READING

- Graphic design – any and all books by Stephen Few and Edward Tufte.
- Color themes – www.colorbrewer.org
- Storytelling – *The Story Factor: Inspiration, Influence, and Persuasion Through the Art of Storytelling*, 2001, by Annette Simmons.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Chuck Pirrello
Enterprise: Product Manager, Visual Analytics, SAS
Work phone: 919-531-4918
E-mail: Charles.Pirrello@sas.com
Web: www.jmp.com