Problems with Bubble Plots

Your computer may be lying to you

Donald J. Wheeler

With the click of your mouse you can turn a list of values into a bubble plot. No thought or effort is required. Simply sit back and let the software gods do the heavy lifting of transforming your list of numbers into a fancy graph. What could possibly go wrong?

In the Dec. 22 issue of ***Science News Magazine*** a bubble plot was used under the headline "Here are all of the impact craters known to survive on Earth." This bubble plot showed 189 circles of various sizes, where each circle represented a known impact crater. These craters ranged in size from about 50 meters in diameter up to 160 kilometers in diameter. On this graph eight craters were identified by name and size. Instead of attempting to deal with all 189 craters, we shall restrict our attention to these eight named craters. They are:

1. the Barringer meteor crater in Arizona (diameter = 1.19 km);
2. the newly found crater under the ice in Greenland (diameter = 31 km);
3. the Chesapeake Bay crater (diameter = 40 km);
4. the Manicouagan crater in Quebec (diameter = 85 km);
5. the Popigai crater in Siberia (diameter = 90 km);
6. the Acraman crater in Australia (diameter = 90 km);
7. the Chicxulub crater in Yucatan (diameter = 150 km); and
8. the Vredefort crater in South Africa (diameter = 160 km).

Thus, the diameter of the Vredefort crater is 134 times larger than the diameter of the Barringer crater. The bubble plot for these eight craters is shown in Figure 1. Does bubble 8 look like it has a diameter that is 134 times the diameter of bubble 1?

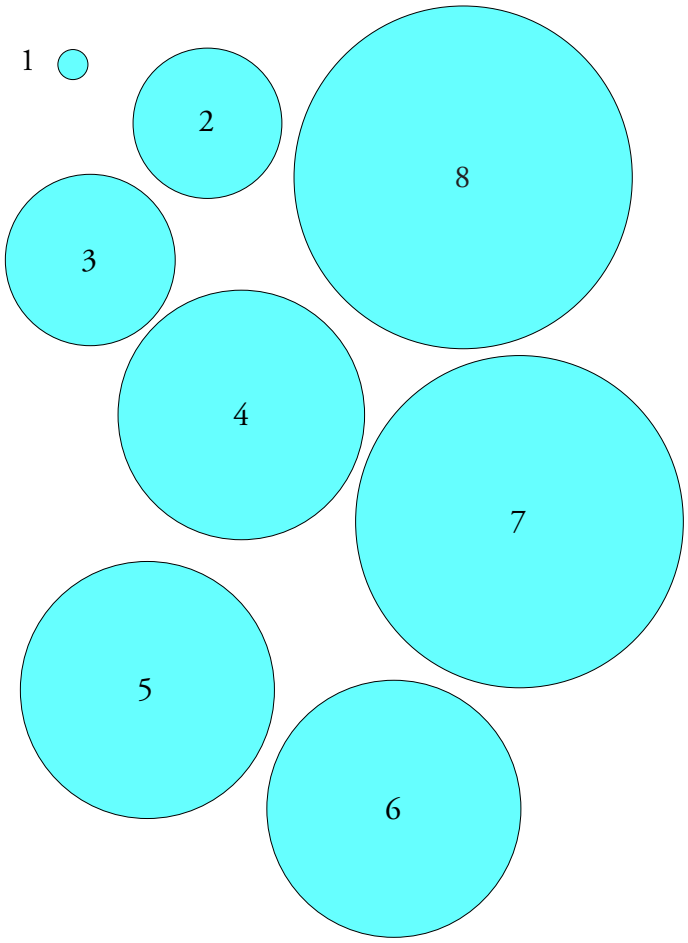

Figure 1: Bubble Plot of Eight Impact Craters

Well, maybe we need to compare the areas. We were all taught that the area of a circle is found by multiplying *pi* times the square of the radius. Using this well known formula the area of the Barringer meteor crater is found to be approximately 1.1 square kilometers, while the area of the Vredefort crater is about 20,100 square kilometers. So, does the largest bubble in Figure 1 look like it is twenty thousand times larger than the smallest bubble?

Figure 2 shows that 121 circles like bubble 1 will essentially cover bubble 8, and that the diameter of bubble 8 is about 11 times the diameter of bubble 1. Clearly neither the relative areas nor the relative diameters of these two craters are correctly shown by the bubbles drawn in Figure 1. Therefore, we must conclude that the bubble plot in Figure 1 does not actually show the relative sizes of the various craters.

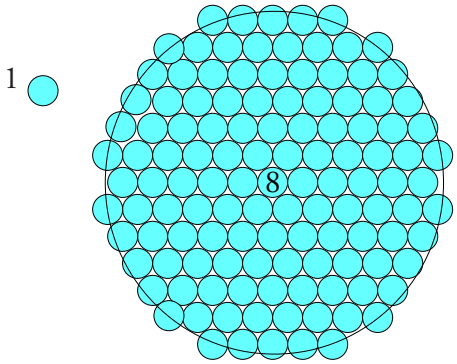

Figure 2: Bubble 8 Covered by 121 Copies of Bubble 1

So what does the bubble plot in Figure 1 show? It attempts to compare the diameters of the craters by using the *area of each bubble* to represent the *diameter of a crater!* By using areas to represent diameters this bubble plot creates confusion. Whenever we use two-dimensional quantities (areas) to represent one-dimensional quantities (diameters) we will inevitably confuse our readers because the variation present in the data will be distorted in the graph.

Another way to illustrate the problem of the bubble plot is to use a bar chart where the bars have the same relative areas as the circles in Figure 1. Since each bar on a bar chart has equal width, this makes the heights of the bars in Figure 3 proportional to the areas of the bubbles in Figure 1. However, to get the same proportions between the bars as between the bubbles we have to use the scale on the left, which is the scale that is *automatically* used by the bubble plot. This automatic scale is equivalent to using a *different linear scale for each bar* of the graph.

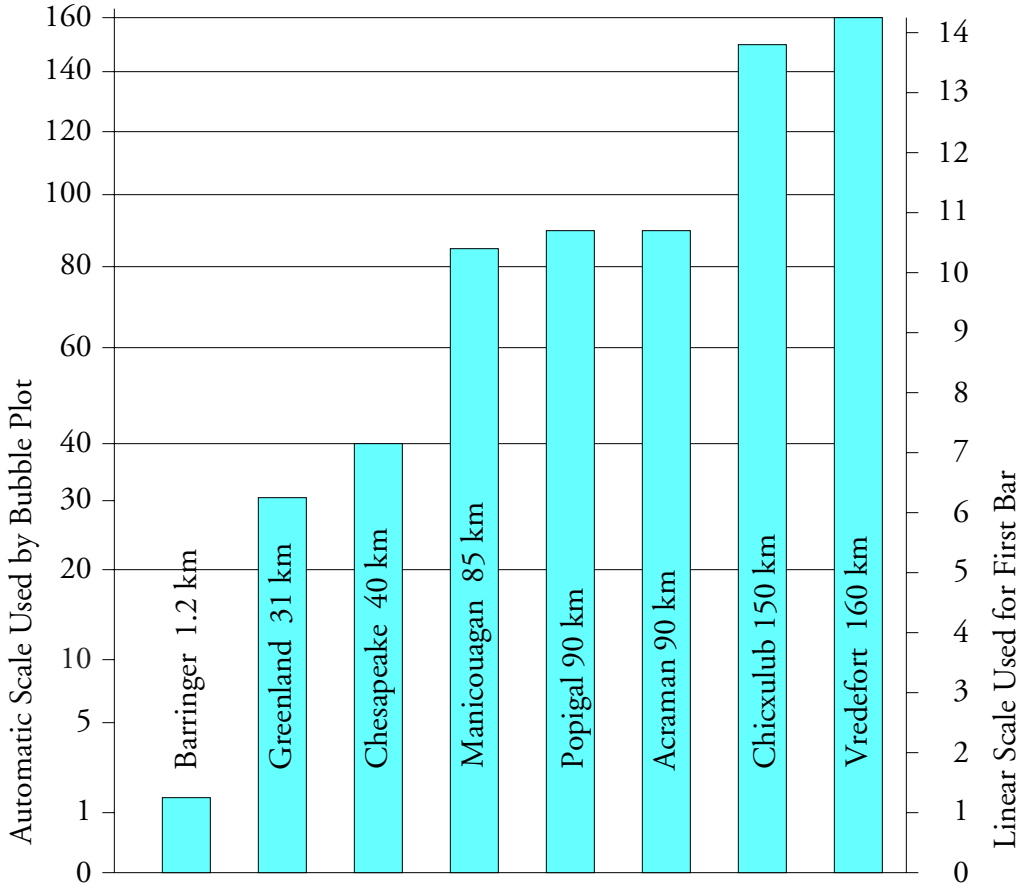

Figure 3: Bar Graph Equivalent to the Bubble Plot

The distortion inherent in all bubble plots may be seen by comparing the linear scale on the right with the nonlinear scale on the left side of the bar chart. Consider the Chesapeake Bay crater and its 40 km diameter.

The scale on the left tells us that 40 km is about 7 times larger than 1.2 km.

The scale on the left also tells us that 40 km is about two-thirds of 85 km.

Finally. the scale on the left tells us that 40 km is also half of 160 km!

Would you want to be the person presenting the graph in Figure 3? You can be—just use a bubble plot!

If we draw a logical graph where the diameters of the craters are actually represented by the diameters of the circles (What a novel idea!) we get Figure 4. There the ratio of the area of the largest circle to the area of the smallest is about 20,000 to 1 as it is supposed to be.

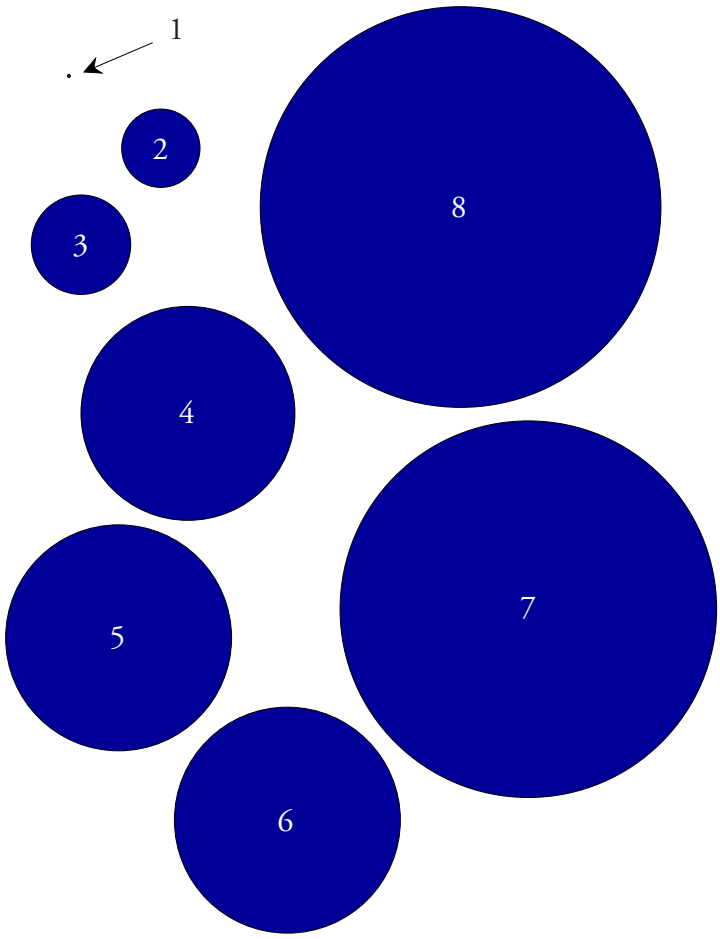

Figure 4: Actual Relative Sizes of Craters

Of course, any graph that displays areas that differ by four orders of magnitude will tend to have elements that threaten to disappear, and this is the case for the Barringer crater in Figure 4. But this is what graphical integrity requires.

Since both Figure 1 and Figure 4 were drawn to have the same total area on the page, we can see the distortion of the bubble plot by superimposing the images for each crater as shown in Figure 5.

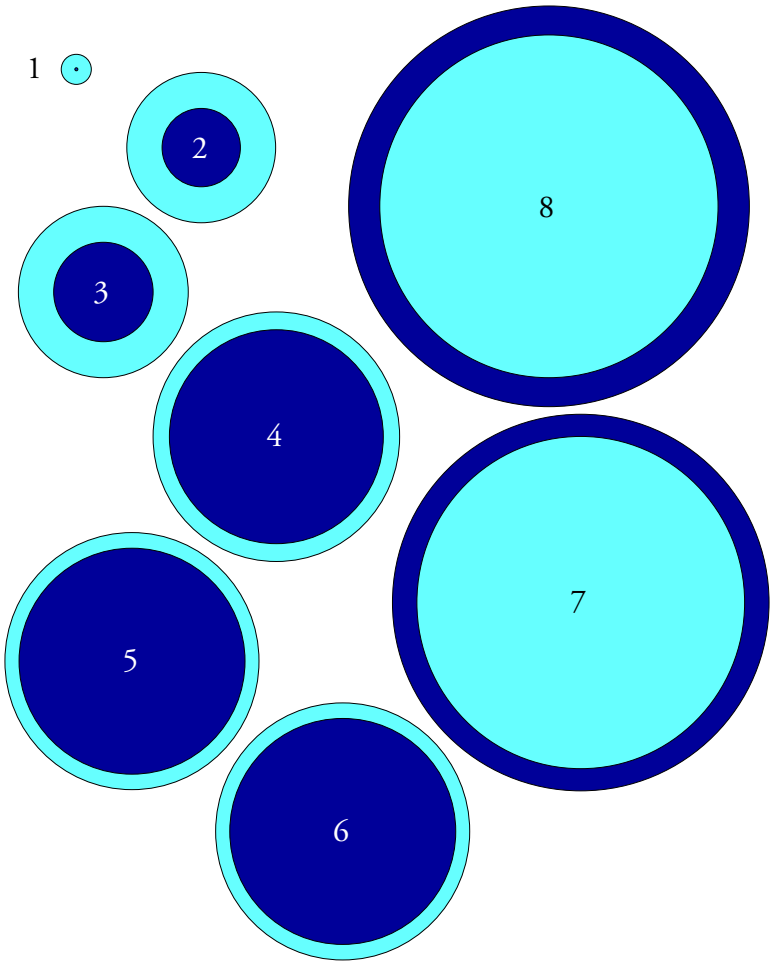

Figure 5: Bubble Plot and Actual Relative Sizes of Craters

The Bubble plot drew the Barringer crater 96 times larger than it should be. It drew the Greenland crater 3.7 times larger than it should be. It drew the Chesapeake Bay crater 2.9 times larger than it should be. It drew the Manicouagan, Popigai, and Acraman craters 1.3 times larger than they should be. And it drew the Chicxulub and Vredefort craters 76 percent and 71 percent as large as they should be. Thus, the bubble plot distorted everything, and as a result it contained virtually nothing that was true, correct, or useful.

GRAPHICAL INTEGRITY

In 1983 Dr. Edward R. Tufte published ***The Visual Display of Quantitative Information***. In this award winning book Tufte warned us about the problem with graphs like bubble plots in the chapter on "Graphical Integrity" when he wrote "Another way to confuse data variation with design variation is to use areas to show one-dimensional data."

In his book Tufte uses the ratio of the variation in the data to the variation shown in the

graph to form a quantity he called the "Lie Factor" of the graph. If we compare the variation in the diameters of the craters with the variation in the diameters of the bubbles, Figure 1 has a Lie Factor of 11.6. If we compare the variation in the areas of the craters with the areas of the bubbles, Figure 1 has a Lie Factor of 134. Either way, the bubble plot in Figure 1 substantially distorts the truth.

In contrast, Figure 4 has a Lie Factor of 1, which means that it does not distort the data. It properly shows the relationships between both the diameters and the areas of the eight craters. It does this by using areas to represent areas, and diameters to represent diameters—it respects the dimensionality of the data being presented.

But wait, doesn't a bar graph use areas to represent one-dimensional quantities?

Not really. While it is true that the bars of a bar graph have area, we force the comparisons to rely upon the heights of the bars by using bars of equal width. The bars have width in order to give them weight so that we can more easily compare the lengths visually. Thus, even though the bars of a bar graph have areas, a bar graph uses a single dimension to represent one-dimensional quantites.

Tufte gives two principles for graphical integrity. The first of these is that "The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities represented." Bubble graphs consistently fail to meet this principle when they use areas to represent one-dimensional quantities.

Tufte's second principle for graphical integrity is that "Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity."

When it comes to presenting data, three simple graphs consistently meet these two principles. They are a simple time-series graph with properly labeled scales, a histogram with equal width intervals, and a bar graph drawn with a complete, unbroken, and linear scale. Decoration and non-data ink (two things that Tufte called "chartjunk") should always be minimized.

So if a stranger, or your software, or even a so-called "friend" suggests using a bubble plot to present your data, just say no. In this way both you and your audience can avoid graphical purgatory.