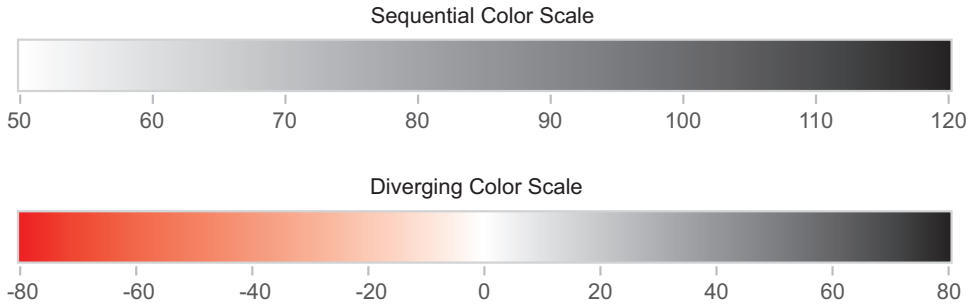
perceptual edge

Heatmaps: to Bin or Not to Bin?

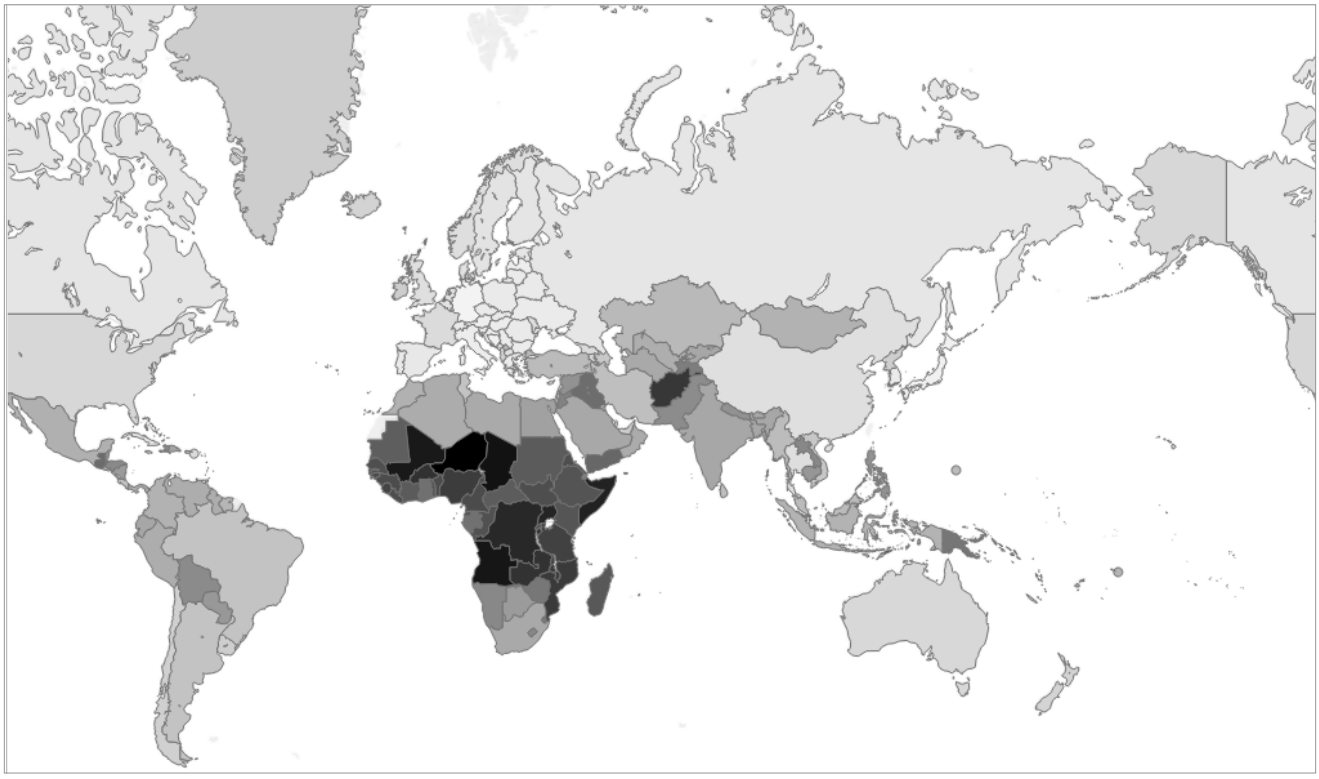
Stephen Few, Perceptual Edge
Visual Business Intelligence Newsletter
October/November/December 2017

Heatmaps have been used to visually represent quantitative data for generations. A heatmap uses variations in color to represent variations in quantity. These variations in color—the color scale—come in two basic types: sequential and diverging. A sequential scale varies in one direction only, from low to high values. A diverging scale varies in two directions, above and below some logical breakpoint.



For example, when displaying profits and losses or temperature in Celsius, zero is a logical breakpoint. Above and below average is another logical breakpoint that can be applied to many measures.

Heatmaps come in several forms. The most common form is a geospatial display called a choropleth map, which fills geographical regions with colors to represent quantitative values associated with those regions.



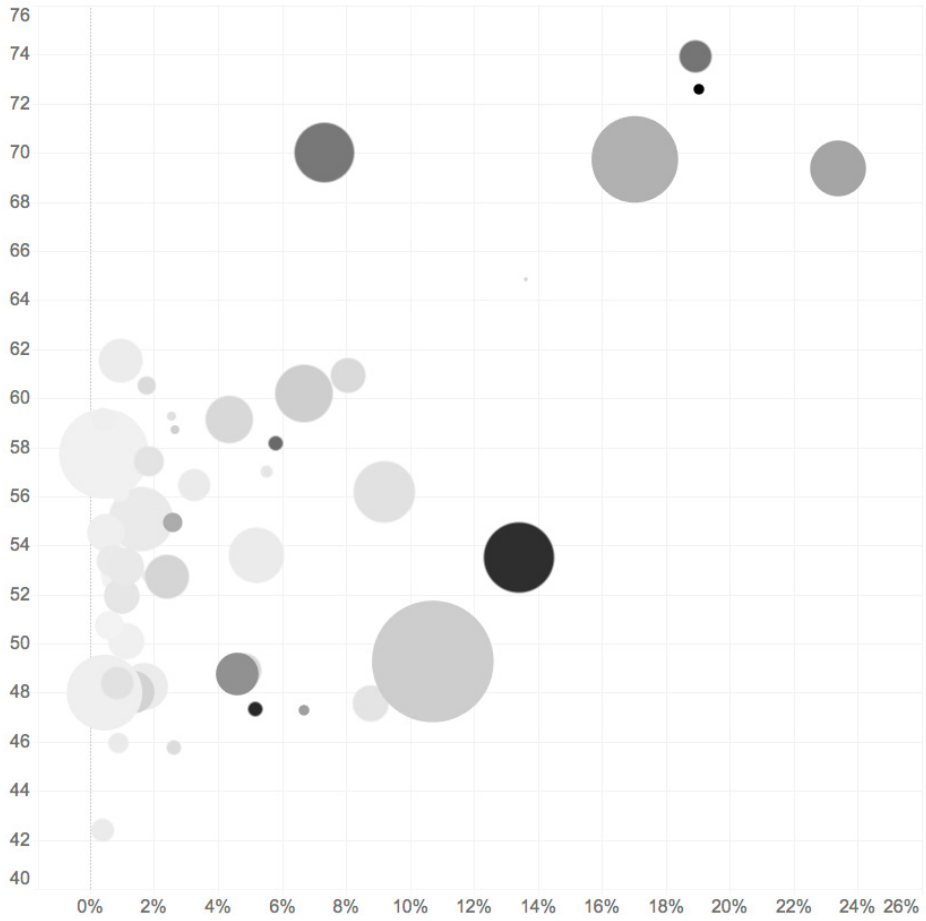
Copyright © 2017 Stephen Few, Perceptual Edge

Another popular form of heatmap, which has become increasingly common in recent years, is the heatmap matrix (a.k.a., heatmap table), which arranges objects—most often rectangles—into columns and rows and fills each of those objects with a color to represent its value.

Room Occupancy Rates

	Room Types				
Hotels	Twin Bed	Queen Bed	King Bed	King Suite	Mean
San Francisco					
La Jolla					
Menlo Park					
San Jose					
San Diego					
Los Angeles					
Oakland					
Berkeley					
Santa Barbara					
Davis					
Redding					
Sacramento					
Fresno					
Palo Alto					
Carlsbad					
Indio					

The heatmap approach to representing values can be applied to data visualizations of many types (dot plots, scatter plots, bubble plots, treemaps, network diagrams, etc.) by coloring objects or sections of objects. In the following example, the heatmap technique has been applied to a bubble plot.

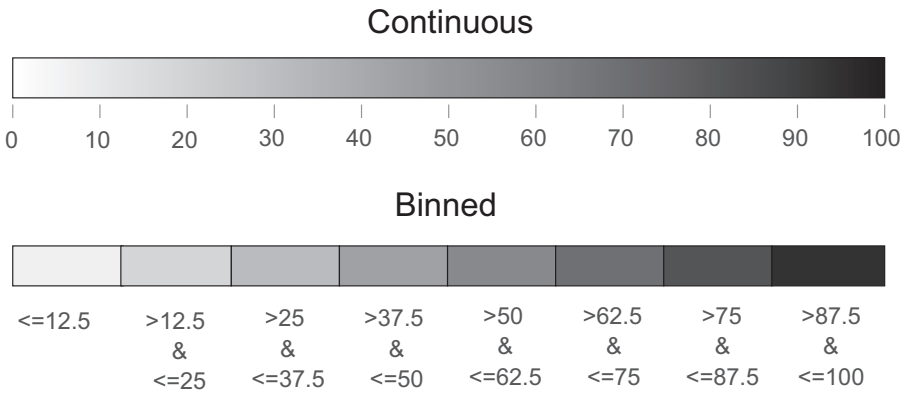


Copyright © 2017 Stephen Few, Perceptual Edge

Variation in color is not ideal for representing quantitative information, for our perception of quantitative differences among colors (i.e., greater or lesser, and to what degree) is not highly developed. However, using visual attributes that we perceive more easily and accurately (e.g., 2-D location and length) is not always an option. When we design heatmaps, however, we should do so in a manner that makes them as perceptually effective as possible.

The Question: to Bin or Not to Bin?

When we create a heatmap, we decide if we will represent quantitative values using a continuous range of colors or one that has been divided into bins of discrete colors.



Both of these scales represent a range in value from 0 to 100. The question that I want us to consider in this article is, “When does it make sense to use bins of discrete colors rather than a continuous range of colors in heatmaps?” In other words, when do colored bins offer advantages over a continuous range of colors? I approach such questions from a particular perspective: data visualizations should reveal all variation in value that is meaningful and useful for the task at hand. To do less reduces the effectiveness of the display.

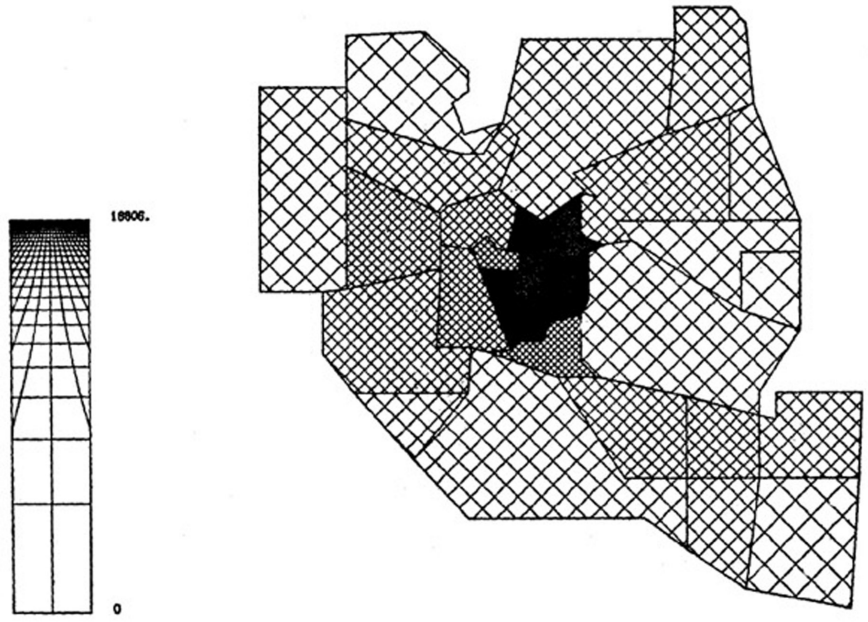
Little of significance has been written about the relative benefits of continuous versus binned color scales. Most of the evaluative work that addresses heatmaps has focused on effective color choices. For example, countless warnings have been issued over the years about the ineffectiveness of a rainbow assortment of colors in heatmaps, because we cannot intuitively perceive an assortment of hues as an ordered sequence of colors. It works best perceptually to vary the range of colors primarily by intensity, varying either 1) lightness, from light to dark, 2) saturation, from pale to bright, or 3) a combination of both, using a single hue or confined set of closely related hues. Best practices for selecting colors is not the topic of this article, however, nor will we consider the question, “Into how many bins should the range of colors be divided?” Instead, we will focus solely on the more fundamental question, “To bin or not to bin?”

The Origins of Binning

If we create a heatmap today using one of the many software tools that support their use, we can usually choose between a continuous or binned range of colors. This wasn’t always the case. Prior to the 1970s, if you created choropleth maps, you always used bins. It would have been difficult, if not impossible, to do otherwise. When assigning colors to the choropleth map manually, which is how it was usually done before the advent of GIS software on personal computers, it wasn’t practical to represent values using a continuous range of colors. Imagine trying to find the corresponding color for each individual value along a continuous range. It was relatively easy, however, to represent values using a few bins of discrete colors. Bins did not emerge because they represented the data more effectively than continuous ranges. Instead, the practice of using bins emerged because it wasn’t practical to do otherwise. Had continuous color scales been technologically feasible in the beginning, the binning technique might not have ever emerged.

Copyright © 2017 Stephen Few, Perceptual Edge

Eventually, in the 1970s, a fellow named Waldo Tobler came up with a way to produce choropleth maps with continuous grayscales using a line plotter, illustrated below.



The legend of pseudo-grayscale colors appears on the left, ranging from low values on the bottom, which appear lightest, to high values on the top, which appear darkest. A sample choropleth map using these pseudo-grayscale colors appears on the right. Tobler's solution was creative, but it doesn't represent a true continuous scale of colors. Nevertheless, his introduction of a continuous range as a replacement for bins was potentially disruptive to cartography.

A heated debate ensued within the cartographic community, with Jean-Claude Muller as the leading advocate of the continuous grayscale method. Michael Dobson led the opposition, expressing several concerns about choropleth maps that used continuous color scales. Essentially, Dobson was concerned that human perception and cognition could not handle the virtually unlimited number of color variations that could be produced by continuous ranges. His concern was ill-founded, however. We can read choropleth maps that use continuous ranges of color quite readily. We live in a world filled with an infinite variation in color. It is natural and hardly overwhelming. This still leaves the question, "Can we perform specific tasks more readily when reading a choropleth map with binned versus continuous colors?" This question has not been reliably answered by empirical research.

Recent Research

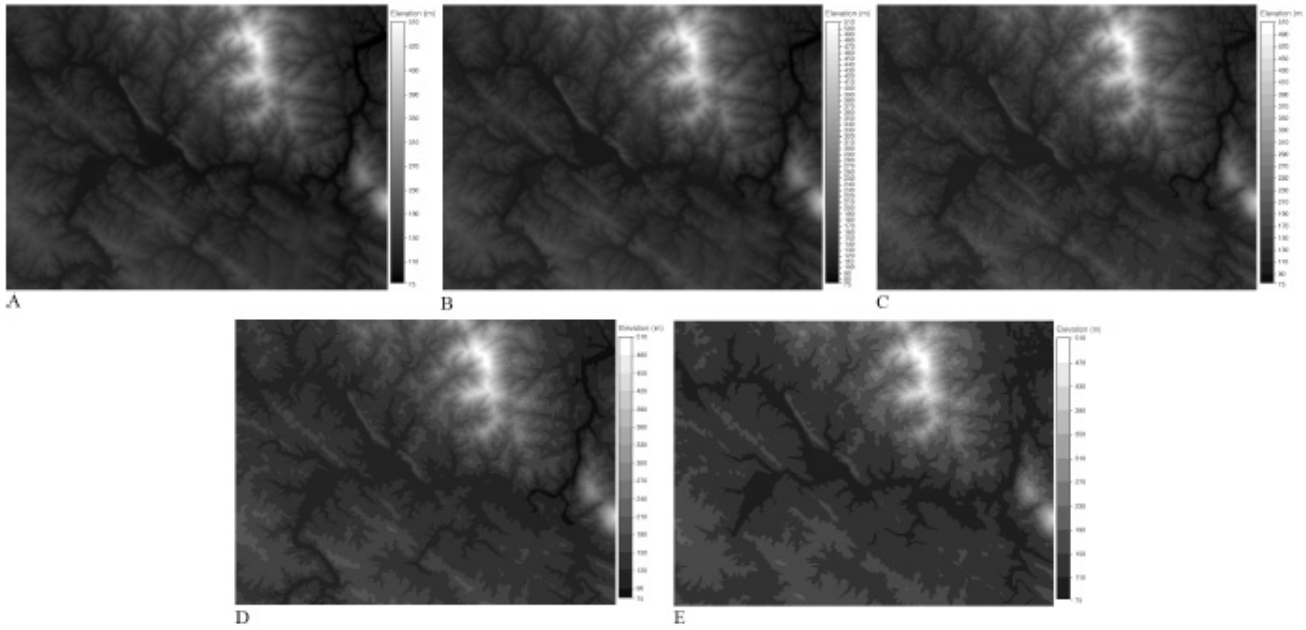
An attempt was made in 2016 by psychology researchers at the University of Utah to determine if particular tasks could be performed more effectively and efficiently when using continuously scaled versus binned heatmaps. Titled "Evaluating the Impact of Binning 2D Scalar Fields" by Lace Padilla, et. al., the paper made a number of observations, including the unexpected finding that test subjects actually performed all of the tasks faster when using continuous scales.

> *The results…show that for some tasks there was no difference in accuracy between binned and continuous encoding (e.g., Compare Heights and Rank Heights tasks), but in other tasks, binned encodings showed higher accuracy than continuous (e.g., Highest Point and Lookup tasks). Interestingly, we found that, generally, participants were faster to complete the tasks with the continuous encoding than with some of the binned encodings. These findings suggest that while participants with continuous encodings appear to complete tasks more quickly, continuous encodings did not improve accuracy.*

Unfortunately, we cannot rely on their findings.

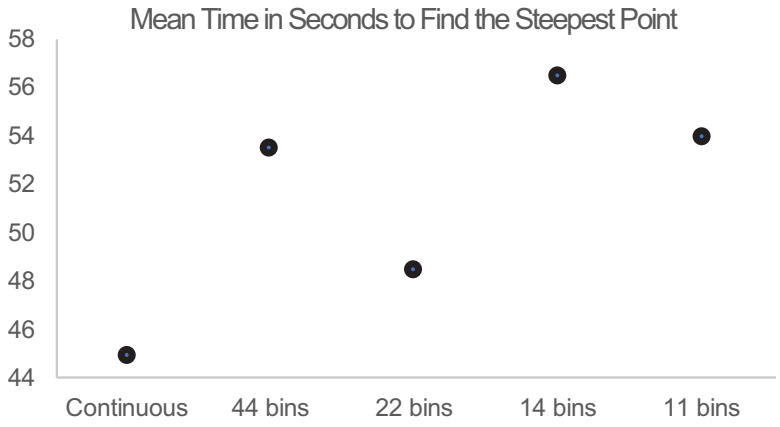
Copyright © 2017 Stephen Few, Perceptual Edge

In the study, in addition to tasks that involved the use of a continuous color scale, separate groups of test subjects performed the same tasks using four different versions of binned ranges consisting of 44 bins, 22 bins, 14 bins, and 11 bins, all of which exceeded the number that is considered practical for choropleth maps, illustrated below.



The design of this study is strange. A sequential scale of binned colors (i.e., one that varies in one direction only from low to high values) is usually limited to a maximum of eight because more than this produces colors that are difficult to discriminate when reading heatmaps. Why the researchers who conducted the "Evaluating the Impact of Binning 2D Scalar Fields" study didn't include choropleth maps with eight or fewer bins to simulate actual practice is a question that they chose not to answer when I questioned them.

The limited glimpses that these researchers provided into their data reveal patterns that appear random. The clearest glimpses are relevant to the time that was required by test subjects to perform various tasks for each of the five conditions (i.e., heatmaps with continuous scales, 44 bins, 22 bines, 14 bins, and 11 bins). Here's the data for one of the tasks:



We would expect specific patterns to emerge, such as a continuous increase in the time moving from continuous scales through 11-bin scales, but no meaningful patterns exist for any of the tasks and no consistent patterns exist among the tasks.

Suspicious data and a few obvious flaws in the design of the study suggests that their findings are unreliable.

Copyright © 2017 Stephen Few, Perceptual Edge

The fact that they have refused to share their data reinforces this judgment. (Note: When I requested a copy of their data, they at first refused altogether, but eventually provided a data file after I wrote an article about this unscientific practice in my blog. They did so, however, in a form that was useless, for the fields of data are cryptically labeled in a manner that renders the data meaningless.) I have a firm recommendation regarding research studies that refuse to provide their data: don't trust them! If we cannot confirm that the data is valid and that it was analyzed in a valid manner, we cannot rely on the results.

Heatmap Tasks

The psychology researchers who produced the study mentioned above and some cartographers, as well, believe that we interact with geospatial versus other forms of heatmaps differently enough that findings related to our performance using one form cannot be applied to other forms. This concern, however, is misplaced, with only a few exceptions. Almost all of the tasks that we perform when reading a choropleth map and other geospatial forms of heatmaps are the same as those that we perform when reading any other form. Here is a list of interactions that I've found to be typical in my own work when reading heatmaps:

- Decode a specific value
- Locate a specific value
- Locate a specific categorical item (e.g., in a choropleth map, the value for a specific country, or, in a heatmap matrix, the value for a specific product)
- Compare two values to determine how and by how much they differ (e.g., which is greater and to what degree)
- Locate the highest or lowest value
- Locate extreme values (e.g., statistical outliers)
- Estimate the average value
- Compare two average values (e.g., averages for each of two regions)
- Rank a set of values
- Identify patterns related to specific categorical items (e.g., Do any particular items in this category, such as particular geographical regions or particular products, exhibit values that are quite different from the overall norm?)

The only exceptions to this sameness of tasks that come to mind are those that are specific to geospatial location, which are few. For example, detecting an association between particular quantitative values and particular geographical features (e.g., major roadways or mountainous regions) is specifically geospatial in nature.

Nevertheless, different arrangements of color-coded objects in heatmaps can affect performance related to perceptual limitations. The two most common examples of this are illusory perceptual effects related to size and adjacency. Large objects of a particular color are more salient than small objects of that same color, which causes us to perceive them differently. This is a problem that plagues choropleth maps and also bubble plots because color-coded geospatial regions and bubbles vary significantly in size. Adjacent colors also affect our perceptions. For example, on a choropleth map, a region of a particular color will appear lighter than another region of that exact same color if it is surrounded by darker regions. We perceive colors relatively rather than absolutely.

Apart from these perceptual problems, which apply to heatmaps of all forms, there is no reason to believe that our performance when examining heatmaps of various forms differs significantly, with one exception, which we'll consider soon.

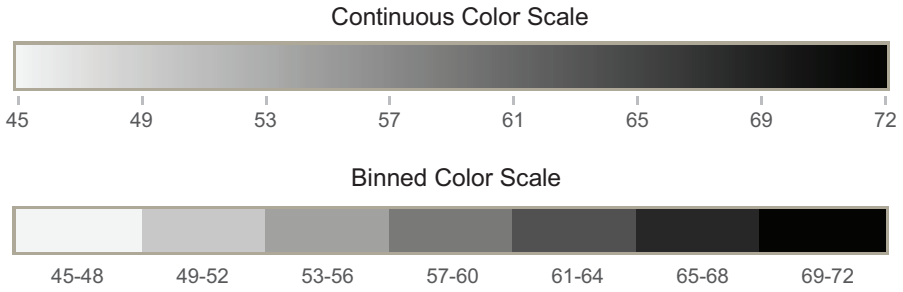
Copyright © 2017 Stephen Few, Perceptual Edge

Back to the Question

Given the fact that binned color scales are a remnant of technological limitations that no longer exist, we should now ask whether we should continue to use them. We should only do so if they offer advantages over a continuous range of colors. Do binned ranges of color offer advantages over a continuous range of colors, and, if so, under what circumstances and for what tasks?

How can we determine what works, what doesn't, and why when no reliable empirical evidence exists on which to base our practices? In such cases, in the absence of reliable experimental evidence, we can rely on our own eyes, our own experience, and our own intelligence.

Binned ranges of color, by their very nature, are less precise than continuous scales. The fewer bins, the less precision, because each bin represents a greater range of values as the same color. Let's consider a typical example. Imagine that we're examining the typical life expectancy of males in Africa. Based on data from 2012, on average male life expectancy in African countries ranged from 45 to 72 years. We could represent these values using a continuous range of color or as colored bins, illustrated below.

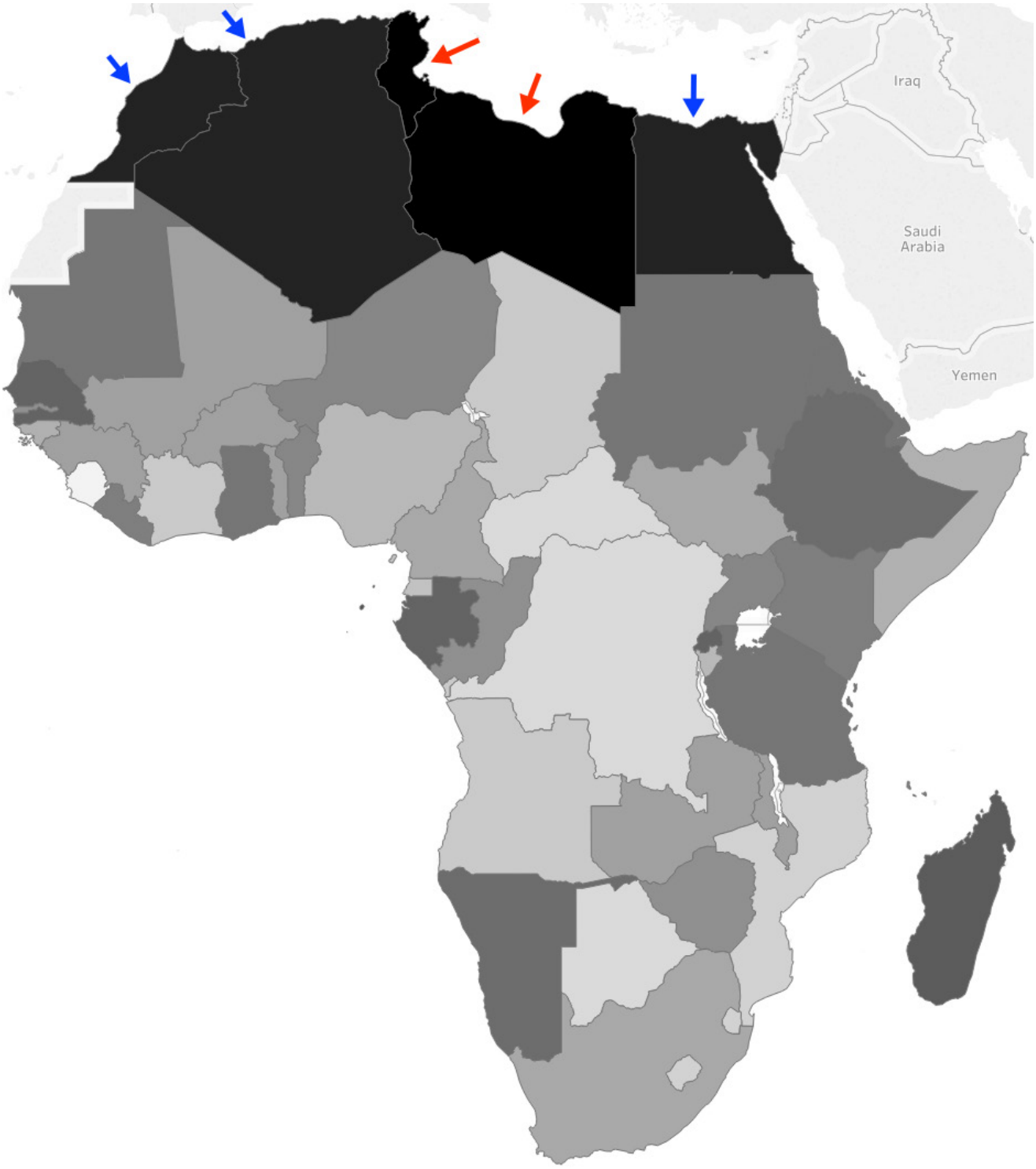


Both scales range in color from light gray on the left to pure black on the right. Using our own eyes reveals the fact that the final two bins in the lower legend are difficult to distinguish, which suggests that seven bins is the practical limit.

In the binned color scale, each bin represents four years of life expectancy. Focus on the first bin on the binned color scale, which represents life expectancies ranging from 45 to 48 years. Now look at that same range on the continuous color scale. We can easily see differences in color between 45 and 48 on the continuous scale which are represented by a single color in the binned scale. The continuous scale represents the values more precisely, but can we perceive those differences when viewing a choropleth map?

Copyright © 2017 Stephen Few, Perceptual Edge

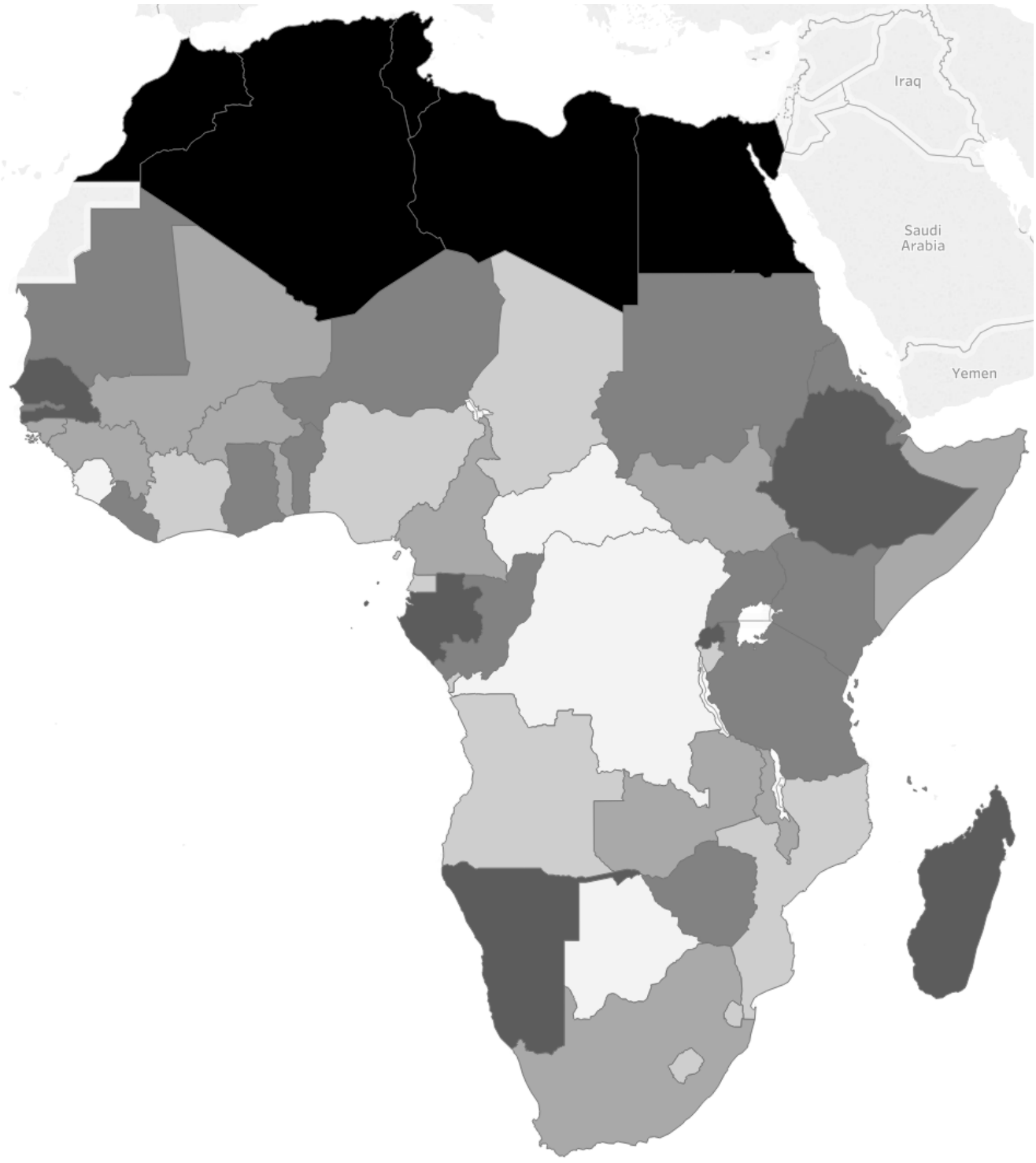
The following map of Africa uses a continuous color scale:



Can you see that the two countries marked with red arrows are darker and thus have higher life expectancies that the three countries marked with blue arrows? If your eyes are as good as mine (my eyes aren't as good as they once were), you can see this, even though the difference is subtle, unless there's a problem with your screen.

Copyright © 2017 Stephen Few, Perceptual Edge

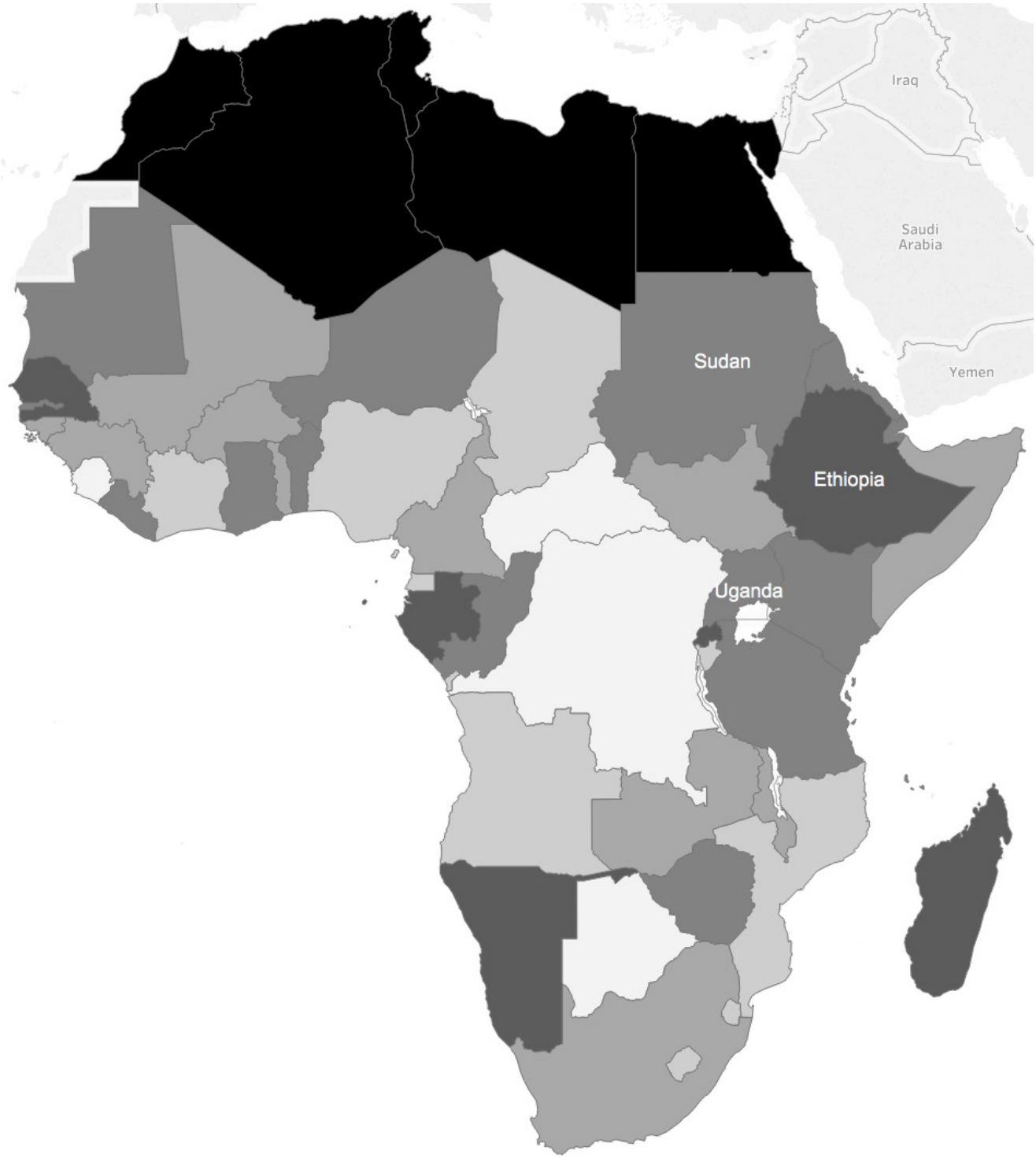
Now look at a version of this same data using a color scales consisting of seven bins:



Can you still see that the two countries previously marked with red arrows are darker than those marked with blue arrows? If you can, you're suffering from an illusion. All five countries at the top of Africa have male life expectancies that are contained within the same 69 to 72 year bin: the two that were previously marked with red arrows have values of 72 years and the other three have values of 69 years.

Copyright © 2017 Stephen Few, Perceptual Edge

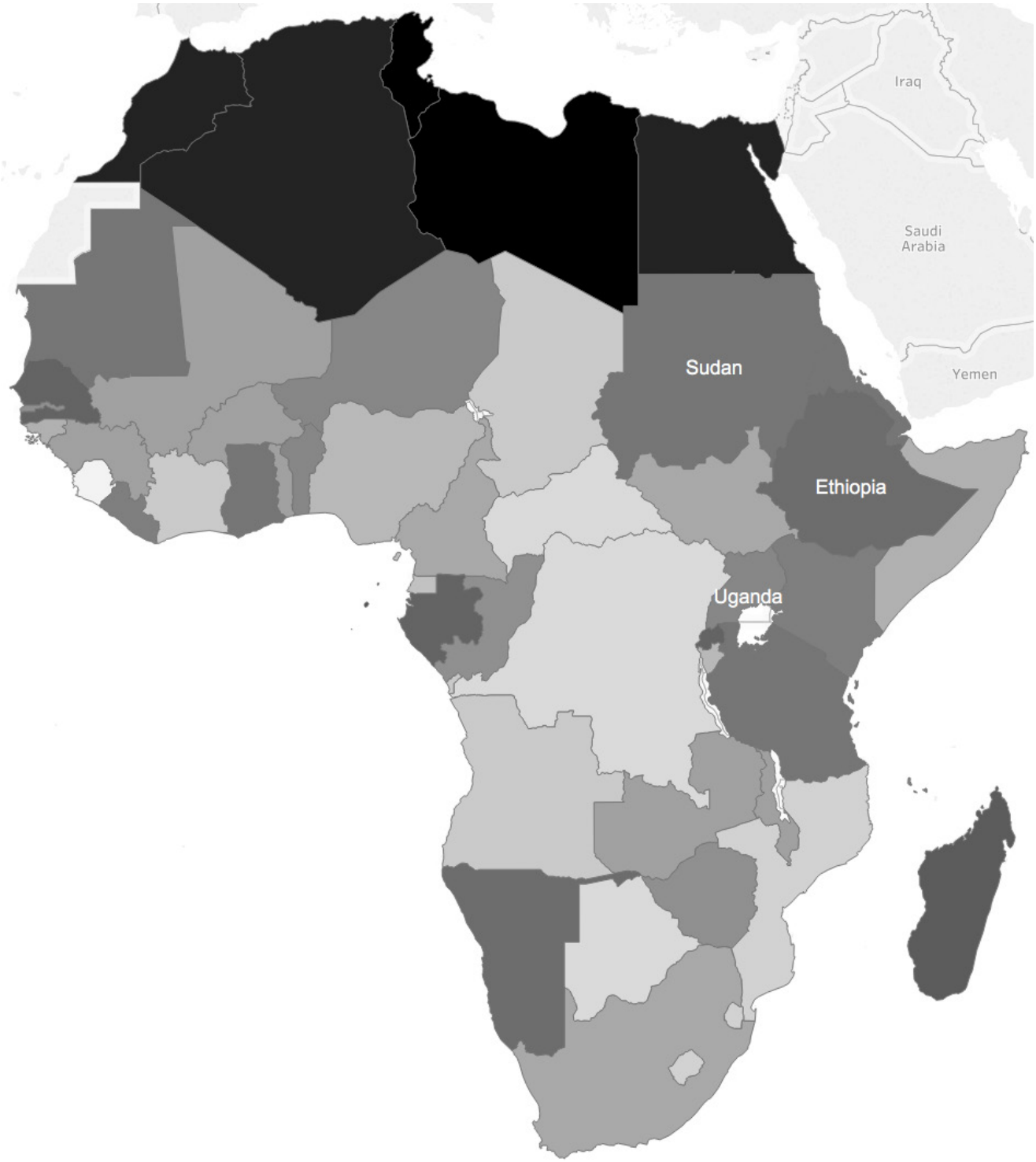
Let me illustrate a similar problem that is built into binned color scales. In the map below with the same 7-bin color scale, I've labeled Sudan, Ethiopia, and Uganda. It appears that the lifespan of males in Ethiopia is significantly greater than in Sudan, and Sudan and Uganda are roughly the same—correct?



Actually, life expectancies in Sudan and Ethiopia are only one year apart even though Ethiopia appears significantly greater (Sudan is 60 years and Ethiopia is 61 years), while Sudan and Uganda differ by two years

Copyright © 2017 Stephen Few, Perceptual Edge

(Sudan is 60 and Uganda is 58), yet they look the same. These differences are accurately displayed using the continuous color scale below.

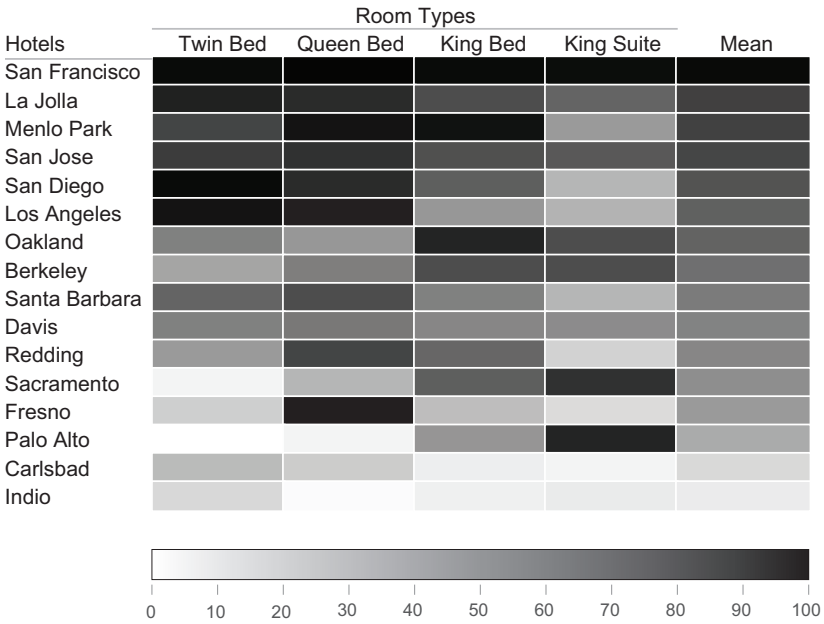


Some advocates of binned ranges argue that the greater precision that a continuous scale provides doesn't matter because we cannot perceive those differences. As you can see, however, this is only true when differences are tiny.

Copyright © 2017 Stephen Few, Perceptual Edge

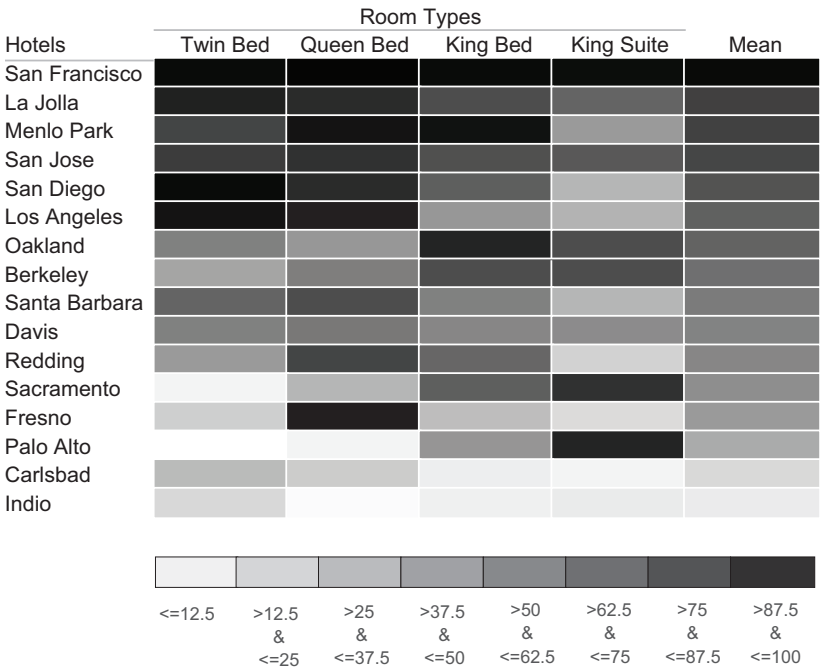
The One Apparent Deficiency of Continuous Color Scales

My own eyes and experience suggests that continuous color scales support all of the tasks that we perform when examining a heatmap with equal or greater effectiveness than binned color scales, with one exception: decoding a specific value. It is difficult to match a specific color that appears in a heatmap with the corresponding color in a continuous color legend unless they appear next to each other. Try it for yourself. In the example below, what is the occupancy rate for King Suites in La Jolla based on the legend, which ranges from 0-100%?



The answer is 65%. Chances are, you were in the ballpark, but off by a fair amount and considerably uncertain.

Assuming that colors are adequately distinct from one another in a binned legend, it is usually easier to match a specific color that appears in a heatmap with the corresponding bin in the legend. Using the binned heatmap below, you can probably match the color representing King Suites in La Jolla with the >62.5 and <=75 bin in the legend, although you were perhaps tempted to associate it with one of the immediately adjacent bins.



Copyright © 2017 Stephen Few, Perceptual Edge

We must keep in mind, however, that each of these bins represents a range of uncertainty that spans 12.5 percentage points. Matching colors in the heatmap to associated bins might be easier, but we might gain little if any advantage in accuracy.

This one potential deficiency of continuous color scales is not significant, however, for we don't (or shouldn't) use heatmaps if decoding specific values is one of our primary tasks. Heatmaps in all forms enable only rough approximations of the values that they encode. Visual perception neither supports precise nor easy quantitative comparisons of colors. When we need precise values, a table of numbers usually works best. When visual representations of the data are more useful than tables of numbers, we can optimize precision by encoding the quantitative values as either 2-D positions or length differences rather than color differences. We should also keep in mind that the difficulty of decoding specific values when using continuous color scales is only a problem when examining printed displays, for computer-based displays can support easy access to precise values whenever they're needed, such as by hovering over specific locations on the heatmap to cause the value to appear as a number.

Bin Rarely

Given the relative benefits of continuous versus binned color scales, it usually makes sense to use bins only when the specific ranges of those bins are meaningful. In other words, we should use bins only when they represent quantitative ranges that represent distinct meanings. For example, in a heatmap that displays the performance of students in a set of assignments, if only the letter grades A (90-100%), B (80-89%), C (70-79%), D (60-69%), and F (0-59%) matter, then it might make sense to use a binned color scale corresponding to these grades. In doing this, we should realize that we're losing information that could be useful, such as huge differences between scores in the F grade range, but we can choose to do this intentionally when circumstances suggest that it's useful.

In all other cases, with possible rare exceptions, continuous color scales represent the data more effectively than binned color scales, so we should stick with them. So, how should we answer the question, "To bin or not to bin?" The informed and conscientious answer is "Rarely." Treat continuous color scales as your default approach to heatmaps, choosing bins only when they offer clear advantages.

Discuss this Article

Share your thoughts about this article by visiting the Heatmaps: to Bin or Not to Bin? thread in our discussion forum.

About the Author

Stephen Few has worked for over 30 years as an IT innovator, consultant, and teacher. Today, as Principal of the consultancy Perceptual Edge, Stephen focuses on data visualization for analyzing and communicating quantitative business information. He provides training and consulting services, writes the quarterly *Visual Business Intelligence Newsletter*, and speaks frequently at conferences. He is the author of four books: *Show Me the Numbers: Designing Tables and Graphs to Enlighten*, Second Edition, *Information Dashboard Design: Displaying Data for at-a-Glance Monitoring*, Second Edition, *Now You See It: Simple Visualization Techniques for Quantitative Analysis*, and *Signal: Understanding What Matters in a World of Noise*. You can learn more about Stephen's work and access an entire library of articles at www.perceptualedge.com. Between articles, you can read Stephen's thoughts on the industry in his blog.

Copyright © 2017 Stephen Few, Perceptual Edge