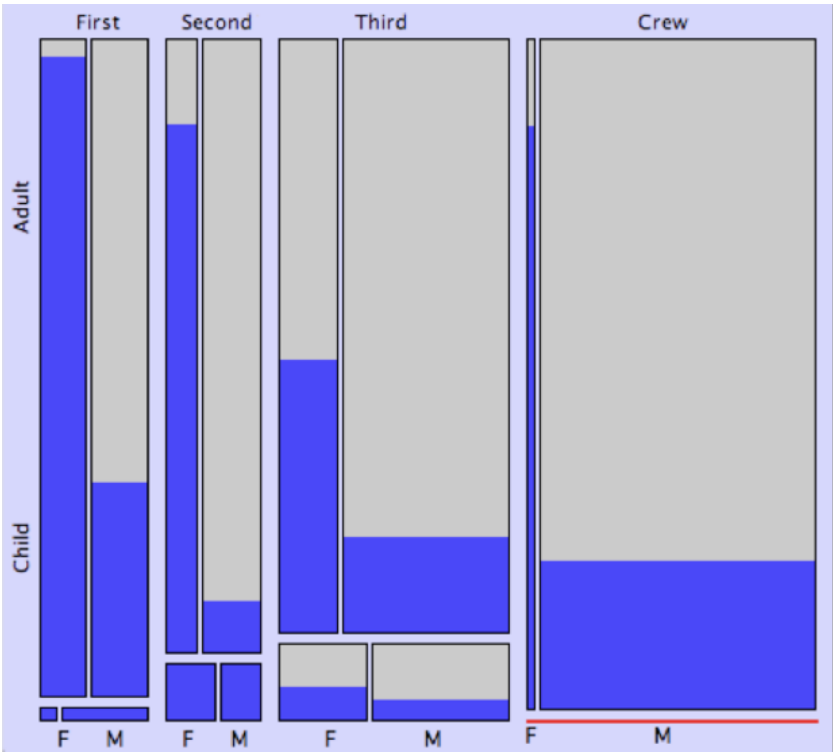
perceptual edge

Are Mosaic Plots Worthwhile?

Stephen Few, Perceptual Edge
Visual Business Intelligence Newsletter
January/February/March 2014

Statisticians are trained in quantitative analysis, yet statistical graphics are often poorly designed. While focusing on the integrity of the data, which I would never discourage, statisticians seldom understand their own perceptual and cognitive abilities while employing graphics to explore and analyze data, nor do they consider the abilities and interests of their audience when presenting their findings. This is especially true regarding the aesthetics of visual design—visualizing data in ways that are pleasing to the eye—which can engage people in the data when done well or turn them away in disinterest (or even disgust) when done poorly. Of course, there are outliers among statisticians: those who take the time to fully respect data by learning and applying the entire gamut of statistical best practices, from the underlying calculations to choices of colors for graphs.

The *mosaic plot* is a graph that is often found in the statistician's tool chest. Below is an often-cited example of a mosaic plot, which tells the story of those who survived when the Titanic hit an iceberg and sunk to the depths.



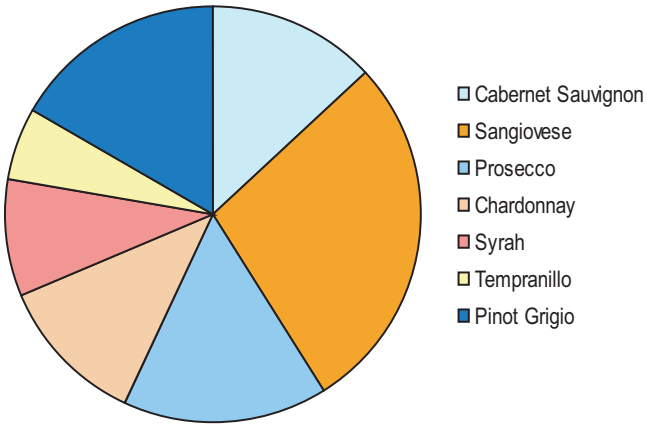
In this example, we see counts of the people who survived (blue) or died (gray) divided into their relative proportions across and within three categorical variables: sex (M for male or F for female), class (First, Second, Third, or Crew), and age (Child or Adult).

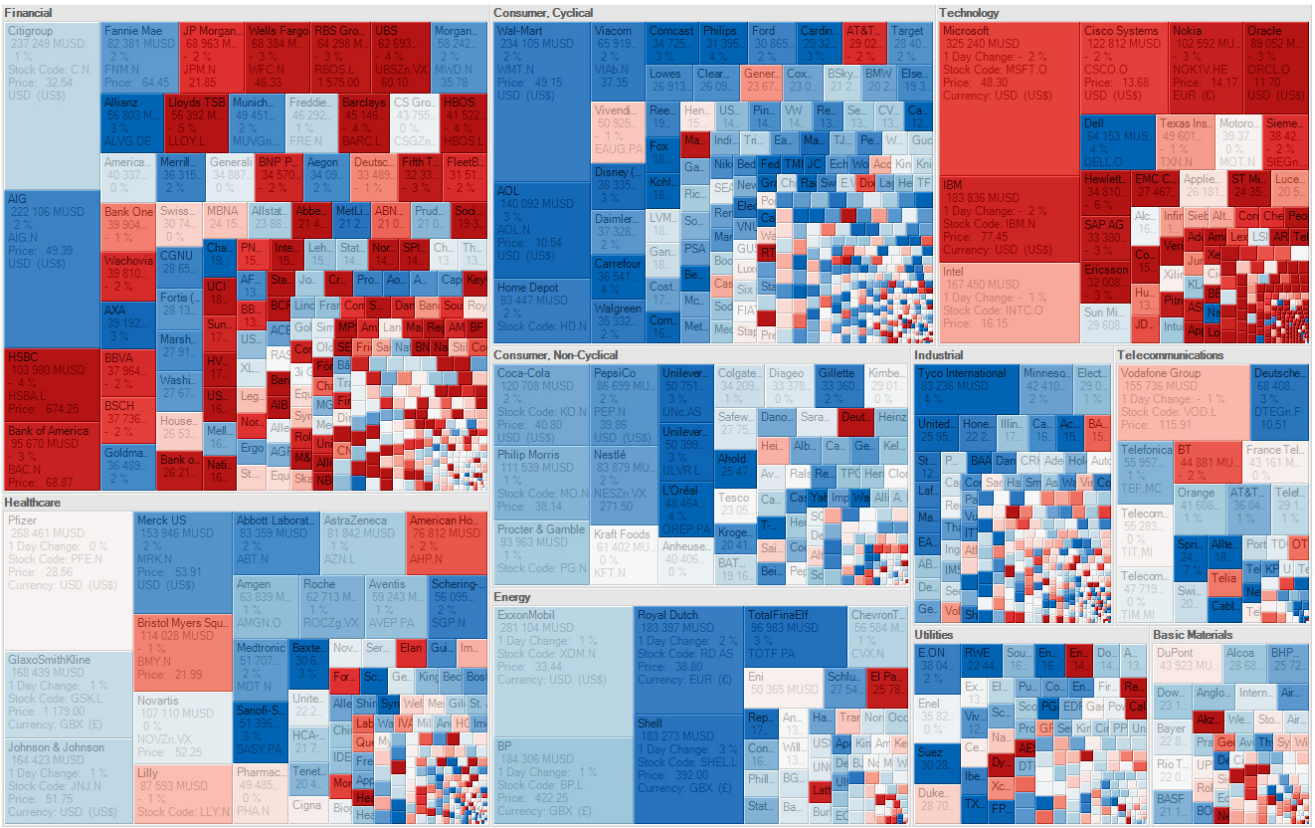
Although they can be found in many statistical software products, mosaic plots are rarely found in analytical products for a general audience, such as business intelligence (BI) products. This could change in time, however, which is why I'm taking the time now to evaluate the merits of mosaic plots, before this happens.

The Occasion

About a year ago I was prompted to take a close look at mosaic plots by a specific event and subsequent discussion. While doing some work for the U.S. Census Bureau, I was invited to participate on a panel with my friend and colleague Naomi Robbins, the author of *Creating More Effective Graphs*, and Richard Heiberger, a professor of statistics at Temple University. Naomi and Richard both made brief presentations about graphical displays of census data and I was asked to comment on them. At one point during Richard's presentation he showed an example of a mosaic plot and I immediately spoke up and said: "I've never seen an example of a mosaic plot that couldn't be presented more effectively using a different approach." Richard politely responded that he had an example that he could show me later, which might change my mind. This led to a series of emails between Richard and me about the merits of mosaic plots.

The mosaic plot resides in the category of visualizations that feature part-to-whole relationships. The best-known graph in this category is a pie chart. A tree map is also a part-to-whole display. Both represent parts of a whole proportionally by means of *containment*. A pie chart is a circular container, which is divided into slices. A tree map is a rectangular container, which is divided into smaller rectangles.





In both cases, an object, which represents the whole, *contains* smaller parts that encode proportions of the whole by their relative sizes. Containment is not the only way to represent parts of a whole. It is a way, however, that is based on an intuitive visual metaphor. When we see an object that is divided into parts, we automatically assume that we're seeing parts of a whole. For part-to-whole displays to work effectively, however, the metaphor alone is not sufficient; we must be able to interpret the values, compare them, and see relationships among them with ease, clarity, and accuracy.

In an article titled "Understanding Area Based Plots: Mosaic Plots" (www.theusrus.de/blog/, September 18, 2011), Martin Theus begins by proclaiming:

> *Mosaic Plots are the swiss army knife of categorical data displays. Whereas bar charts are stuck in their univariate limits, mosaic plots and their variants open up the powerful visualization of multivariate categorical data.*

Let's find out if Theus' enthusiasm is justified.

Introducing the Mosaic Plot

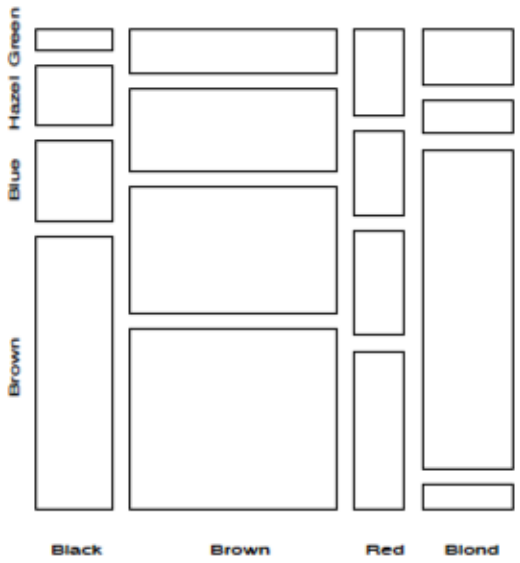
In Michael Friendly's paper, "A Brief History of Mosaic Plots" (2001), he provides the following description:

> *In statistical graphics, the mosaic display, attributed to Hartigan and Kleiner (1981), is a graphical method to show the values (cell frequencies) in a contingency table cross-classified by one or more "factors".*

A *contingency table* is simply a table that displays a count (frequency) in each cell that resides at the column and row intersections of two or more categorical variables. Consider a group of individuals for whom data was collected regarding two variables: hair color (black, brown, red, and blond) and eye color (brown, blue, hazel, and green). (This example comes from Friendly's paper.) Below is a sample two-way (as in two categorical variables) contingency table of this data:

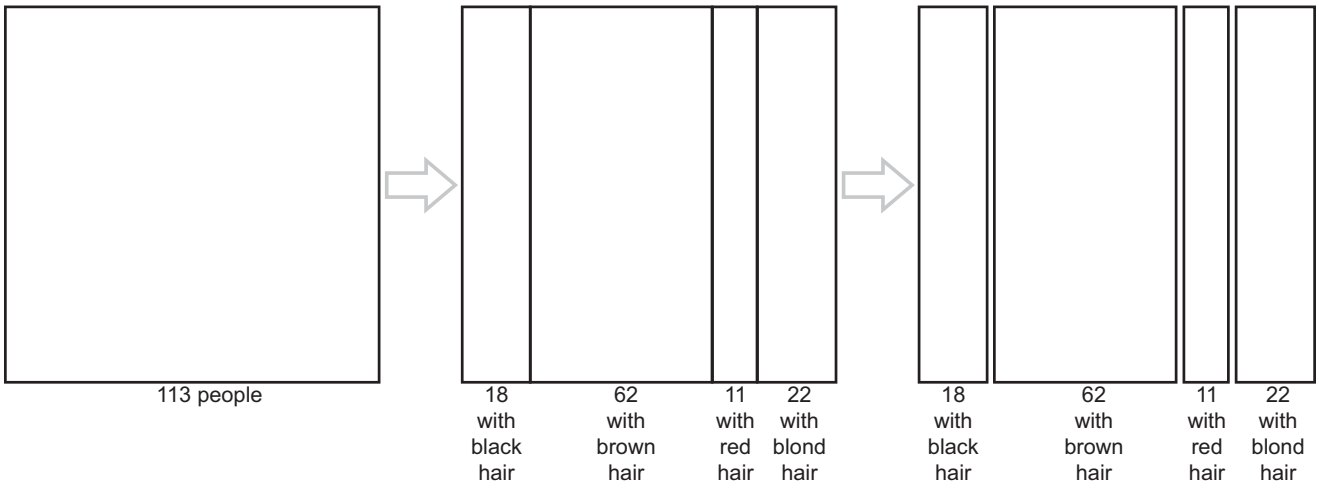
	Hair Color				
Eye Color	Black	Brown	Red	Blond	Total
Brown	10	25	4	2	41
Blue	4	18	3	15	40
Hazel	3	13	2	2	20
Green	1	6	2	3	12
Total	18	62	11	22	113

Here's the same information displayed as a mosaic plot. (Note: The counts in the table above are my approximations of the frequencies shown in Friendly's example below.)

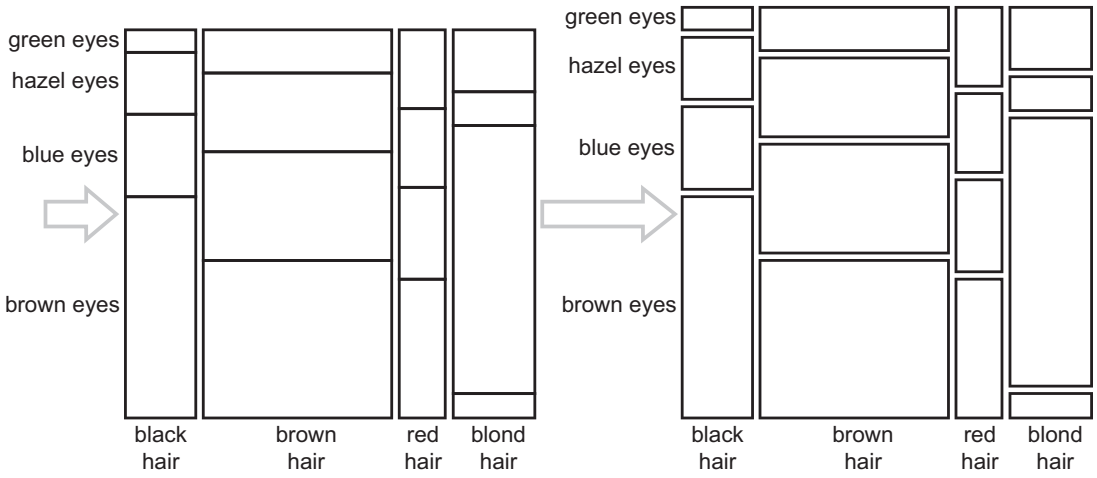


The widths of the rectangles represent the proportion of people with each hair color and their heights represent the proportion of people with each eye color within each hair color group. The area of each rectangle is proportional to the frequency of each combined eye color and hair color group. In other words, the areas represent the numbers in the body of the contingency table.

To create a mosaic plot, you begin with a large rectangle and divide it into vertical sections based on the first categorical variable, in this case hair color, and then you add a little space between the sections.



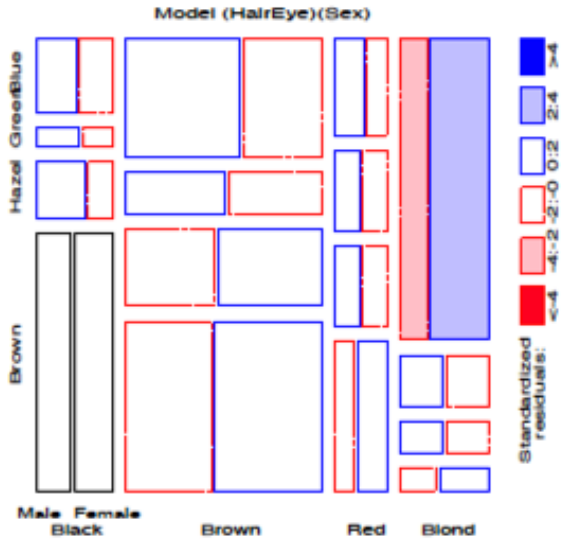
You then divide it into horizontal sections based on the second variable, in this case eye color, and once again add some space between them.



Spaces between the sections are conventional, but not necessary. When these spaces are omitted, the graph is sometimes called a *Mondrian diagram*.

Mosaic plots are not limited to two categorical variables. For example, we could add sex as a third variable by dividing the rectangles that we already have horizontally again to create the following three-way mosaic plot.

(Please ignore the fourth variable that Friendly encoded in this example using colors. We're going to keep things simple.)

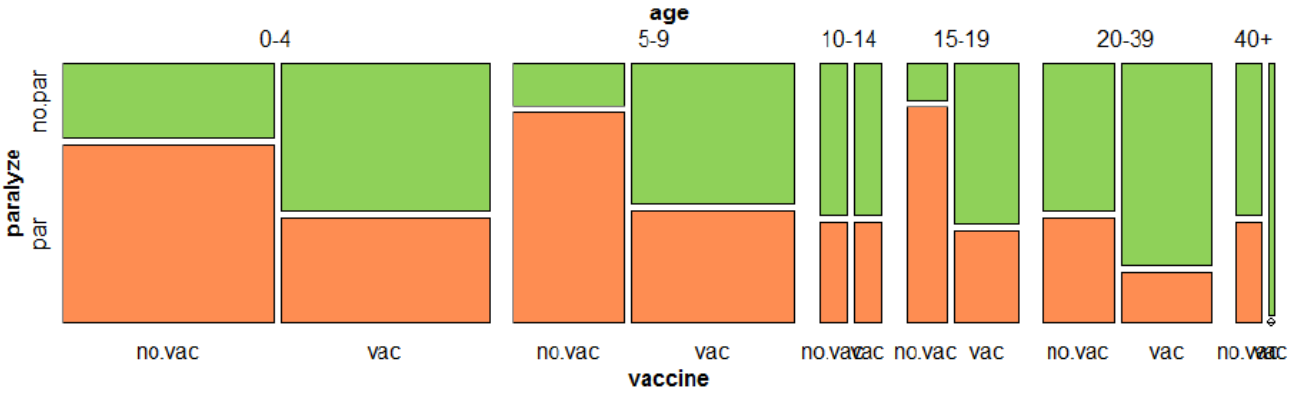


Even more variables could be added, each time by subdividing the existing rectangles, either horizontally or vertically, in alternating order.

As you can see, the concept is fairly straightforward. Let's see how well mosaic plots work for our brains.

Problems with Mosaic Plots

Remember that Richard Heiberger said he could show me an example of a mosaic plot that might change my mind about their effectiveness. Here's the example that he had in mind:



	0-4		5-9		10-14		15-19		20-39		40+	
	no	vac	no	vac	no	vac	no	vac	no	vac	no	vac
no.par	10	20	3	15	3	3	1	7	7	12	3	1
par	24	14	15	12	2	2	6	4	5	3	2	0

Figure 12: Mantel–Haenszel–Cochran test for the Salk polio example. It is easy to see from the mosaic plot that the upper right box in each age group is taller than the upper left box in its own age group. That is, the proportion of cases without paralysis in the vaccinated treatment is higher for all age groups.

In this example, we have a mosaic plot and the three-way contingency table on which it is based. Its three categorical variables are Salk polio vaccine (either administered to the patient or not), paralysis (either happened to the patient or not), and age of the patient in years (0-4, 5-9, 10-14, 15-19, 20-39, and 40+). The colors are there to merely reinforce the distinction between paralysis (orange) and no paralysis (green).

Richard described the primary purpose of the mosaic plot as follows:

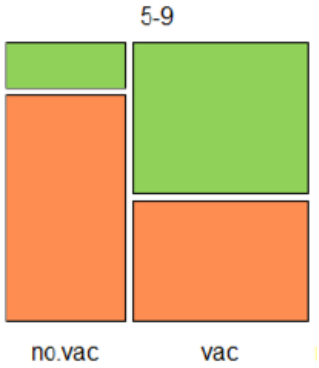
> *We wish to learn if symptom status (paralysis or not) is independent of vaccination status after controlling for age… The goal of the study is to show that vaccination reduces paralysis.*

Richard was using this particular example not so much to promote the merits of mosaic plots as general tools of analysis but as the means to communicate a particular message. Whether it succeeds or fails at this task tells us little about the usefulness of mosaic plots in general. To evaluate the effectiveness of this mosaic plot in light of Richard's stated purpose, we should not merely ask "Does it communicate the message?" but instead "How easily, clearly, and accurately does it communicate this message?" In the next section of this paper I'll show that a specific message such as this can be displayed more simply and clearly using a different type of graph—one that is quite familiar.

Let's not limit our assessment of mosaic plots to their ability to tell specific stories. Let's consider their worth for data exploration, analysis, and communication in general. The fundamental problem from which all graphs that work on the principle of containment suffer is that they encode values as the 2-D areas of objects, which our brains are not well designed to interpret or compare. We accept this difficulty as a worthwhile compromise when we use tree maps, because they enable us to do something that we can't do using more effective means of display: to view and compare huge numbers of values; far too many for a bar graph. The values in pie charts, however, can be displayed in other ways that our brains can interpret and compare with greater ease, clarity, and accuracy. Perhaps this is true of mosaic plots as well. We'll look into this, but first, let's consider some of the other problems that exist in mosaic plots, all of which can be seen in the example that Richard provided on the previous page.

1. Comparisons of rectangle sizes are complicated by the fact that the rectangles can vary dramatically in aspect ratio.

   Which is bigger in the example below (and don't cheat by looking up the answer in the contingency table): the number of children who received no vaccine but suffered from paralysis (the left-hand orange rectangle) or the number of children who received the vaccine and did not suffer from paralysis (the right-hand green rectangle)?



   The correct answer is that neither is bigger for they both represent the same value: 15 children.

2. One variable is encoded as the height of rectangles and the other is encoded as their widths, but it is difficult to focus independently on either heights or widths when they both vary.

   Richard believes this is an advantage of mosaic plots:

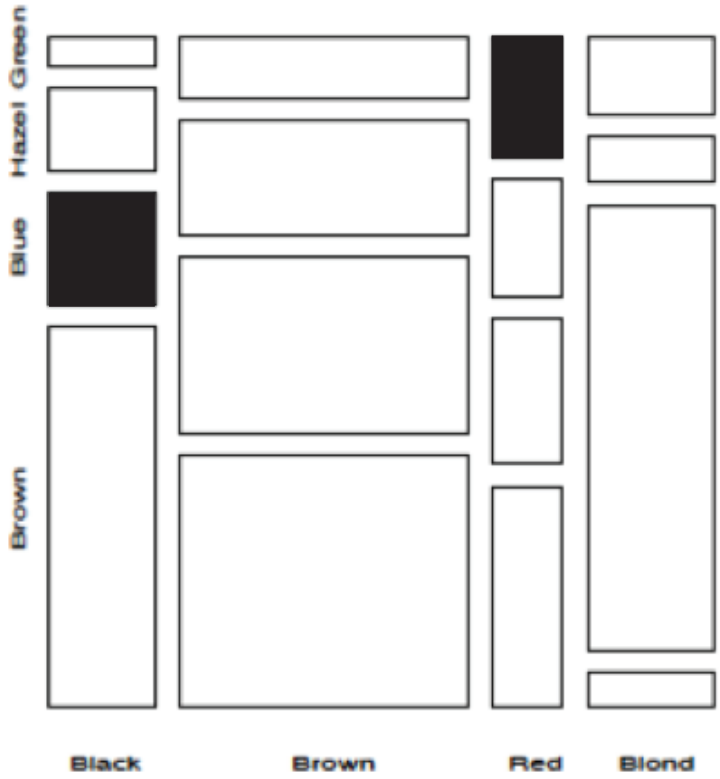
   > *The strength of this form of display is that we do not need to shift our perception. When we compare proportions within each group, we study only height. When we compare counts between groups, we study only width.*

   Contrary to Richard's belief, this is actually a disadvantage, for we must in fact shift our perception. An object's height and width are *integral attributes*, which means that we naturally perceive them

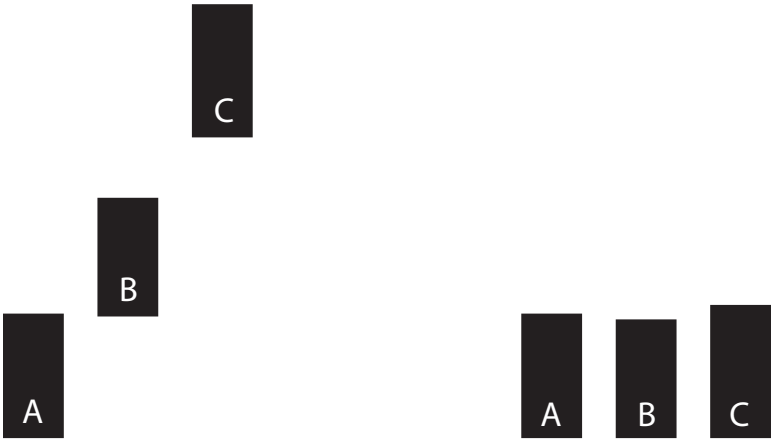
together as a whole (i.e., area) rather than independently. To perceive either independently, we must concentrate on one while actively trying to ignore the other. If you doubt this, look at the mosaic plot again and try to focus on either height or width alone while ignoring the other dimension.

3. It is difficult to compare lengths or heights that are not arranged side by side along a common baseline.

   For example, in the graph below, because the two highlighted rectangles are not aligned along a common baseline, it is more difficult to compare their heights than it would be if they were aligned along a common baseline.

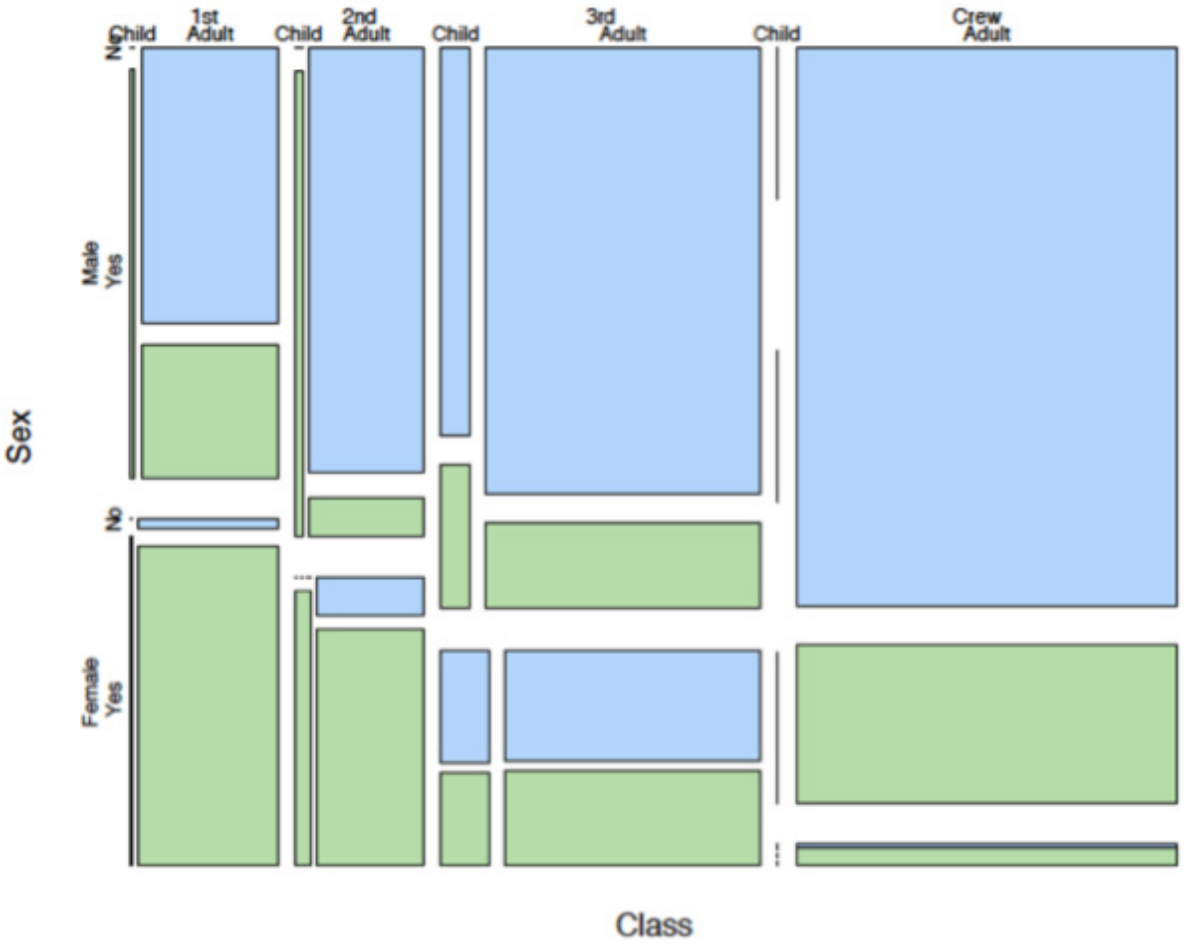


The example below illustrates this problem more clearly. It is easy to compare the heights of the bars on the right and put them in ranked order because they are side by side and share a common baseline. Doing this for the same bars on the left, however, would be difficult.



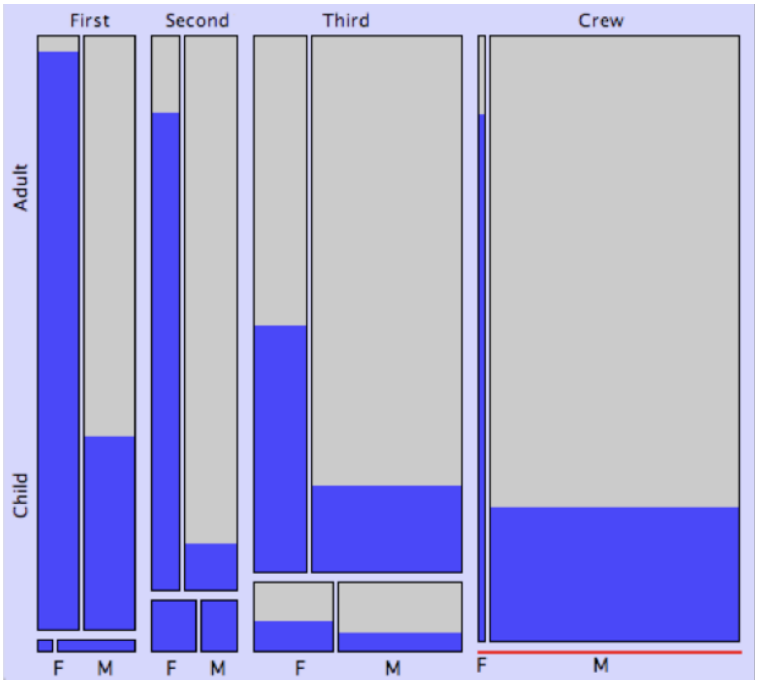
4. It is often difficult to label categorical items.

   Notice that while examining the following mosaic plot about the Titanic disaster below, it would be easy to forget what the various rectangles represent because labels are not clearly associated with most of them.



   Notice also that in the following excerpt from the polio vaccine mosaic plot the labels overlap, which is a common problem.

And in the example below, notice that the Child label on the left appears next to Adult data. Data regarding children in First Class appears as a tiny sliver at the bottom. It's likely that this particular labeling problem was the fault of the software rather than the person using it.



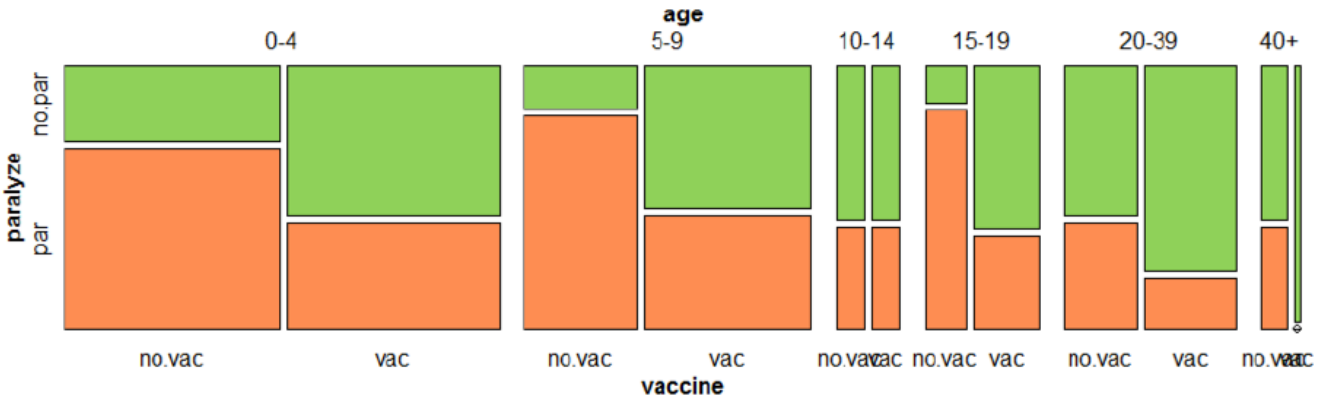
5. Software products sometimes represent data in ways that cause the parts to inaccurately represent the whole.

   Notice that the Titanic example above includes a red rectangle on the bottom right to represent a value of zero, which robs height from the Crew rectangles above it, thereby misrepresenting their values.

In their attempt to put everything into a single container, no matter how complex the data, mosaic plots compromise perceptibility and they do so unnecessarily. Anything that can be displayed in a mosaic plot can be better displayed in one or more bar graphs. This may not be sexy, but as you'll see, it works.

Other Ways to Display Multidimensional Parts of a Whole

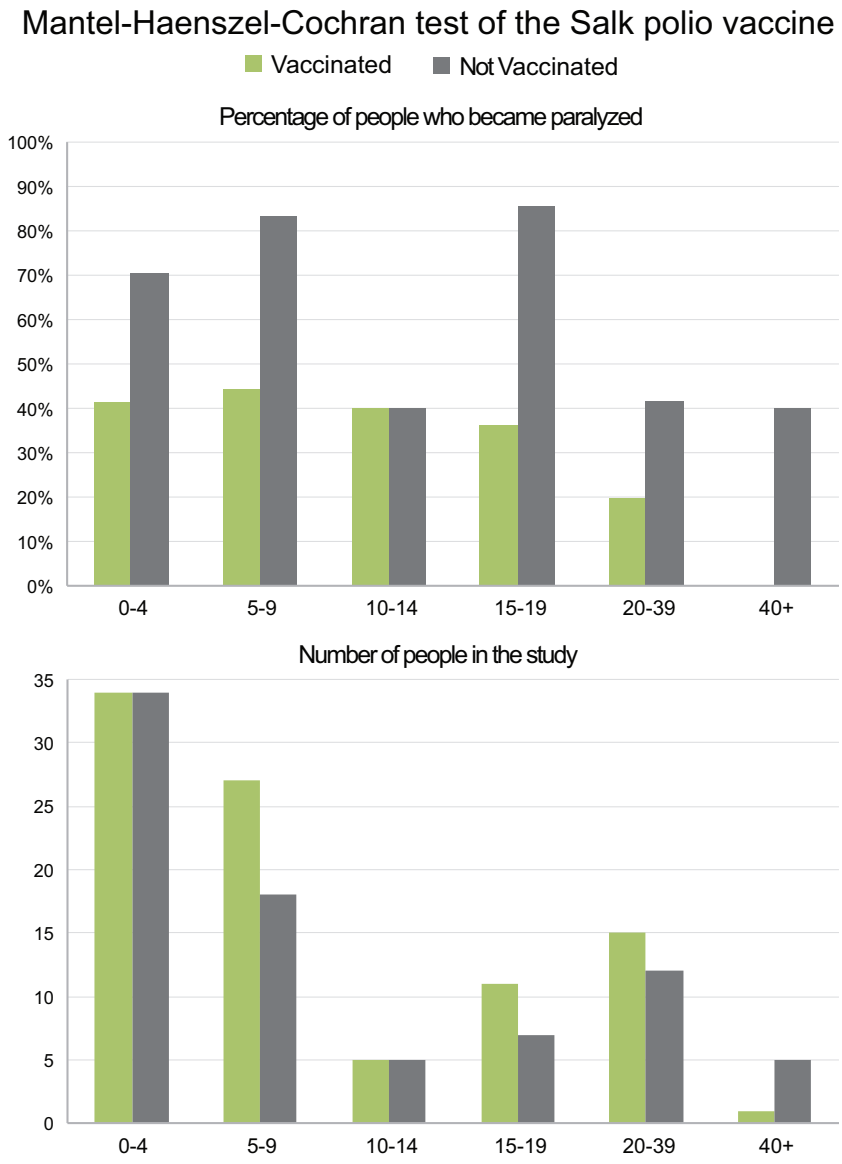
Let's begin by seeing if we can display the Mantel-Haenszel-Cochran test for the Salk polio vaccine more effectively. Here's the mosaic plot again:



Before reading on, take a moment to study this graph. Imagine yourself using it to explore and analyze the results of this study. What's the full story?

When I looked at this mosaic plot initially, what first caught my eye was that some age groups were much wider than others, with 0-4 and 5-9 being the widest and thus largest groups of people in the study. Next, I noticed that the total amount of green (no paralysis) versus orange (paralysis) appears to be roughly equal. With a bit more time, I also noticed that in every age group but 10-14 orange dominated the left side (no vaccine) and green dominated the right side (vaccine). This revealed that paralysis was the likely outcome of not being vaccinated and the opposite was the likely outcome of being vaccinated, suggesting that the vaccine was useful. In other words, I was able to see what Richard intended: when controlling for age, the vaccine seems to reduce paralysis. So, despite the problems that we've found in mosaic plots, this one serves its primary purpose. The question remains, however: "Do mosaic plots provide a necessary and best means to discern stories of this nature?"

Here's the same data displayed as two bar graphs:



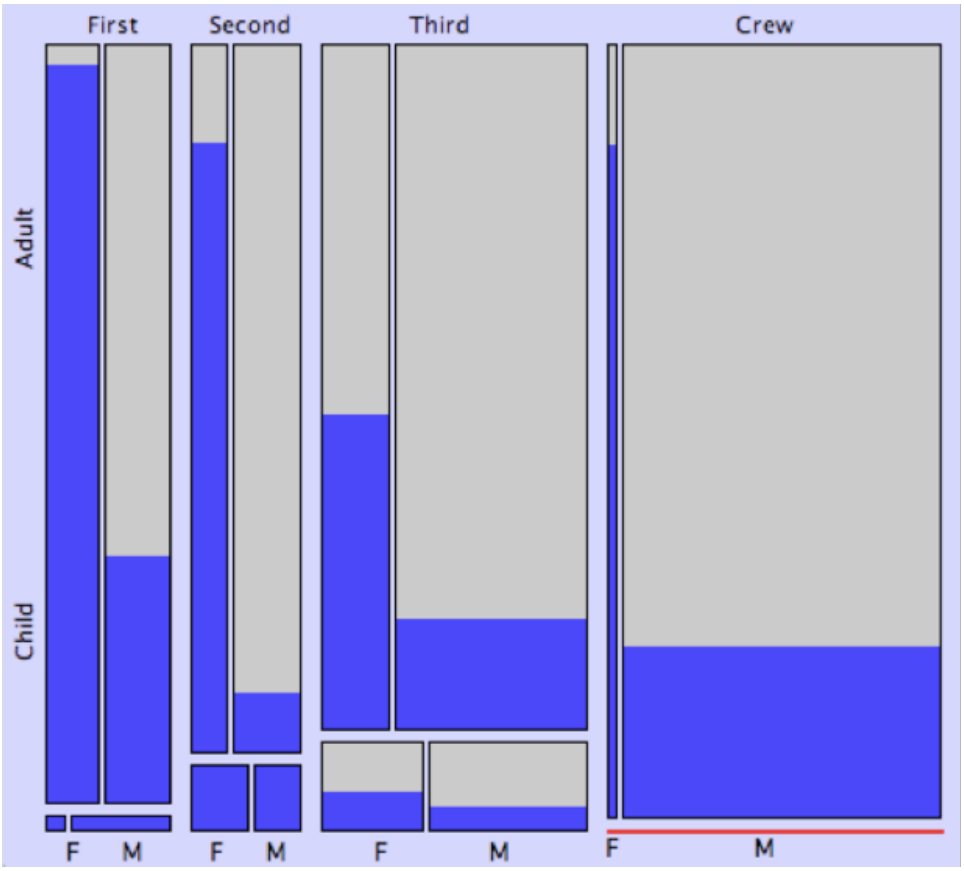
The top bar graph alone tells the primary story of the test results. The gray bars (not vaccinated) that dominate the upper graph of people who were paralyzed show us immediately and clearly that paralysis was reduced by the vaccine. I included the bottom graph primarily to show that the equal percentages of people who were paralyzed in the 10-14 age group whether vaccinated or not could be the result of a small ten-person sample.

Here are three additional facts, among several more, that can be observed in this data set:

- In the 40+ age group, no one who received the vaccine was paralyzed.
- The number of people who participated in this test per age group in rank order from greatest to least are 0-4, 5-9, 20-39, 15-19, 10-14, and 40+.
- The small number of children in the 10-14 year age group stands out as an anomaly, raising the question, “Why was this age group smaller than those that precede (5-9) and follow it (15-19)?”

All of these facts can be seen both in the bar graphs and the mosaic plot, but in no case does the mosaic plot provide a superior view, while the bar graphs are superior in several respects.

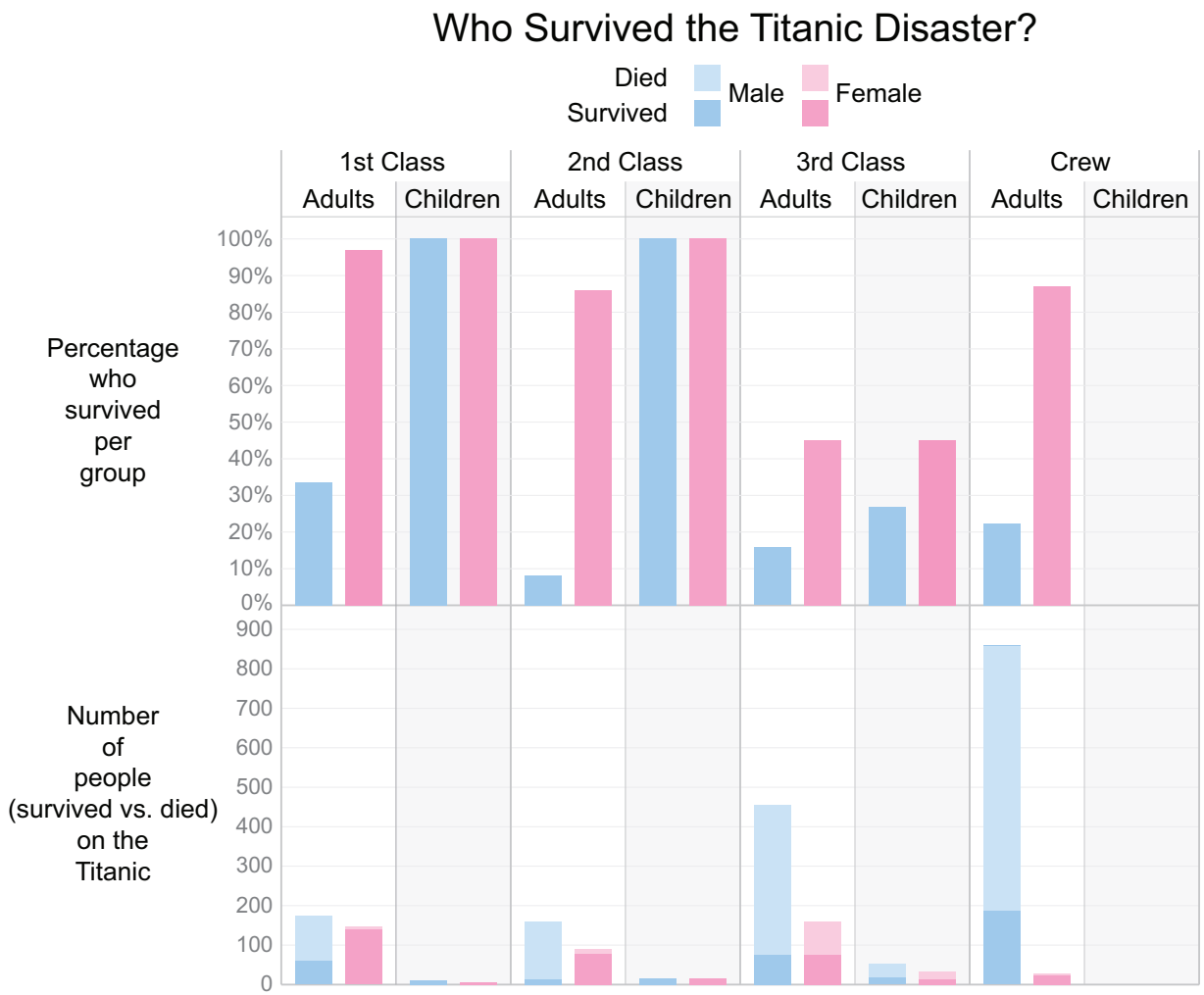
Let’s put bar graphs to the test again using the Titanic data. Here’s how it looks in a mosaic plot:



In this example, survivors are highlighted in blue. Take some time on your own initially to learn the story of those who survived vs. those who died. When you’ve learned what you can from the mosaic plot, continue to examine it by answering the following questions:

- What is the relationship between class and death?
- Did a greater proportion of men or women survive?
- Among the crew, did a greater number of men or women survive?
- What was the relationship between class and the number of children?

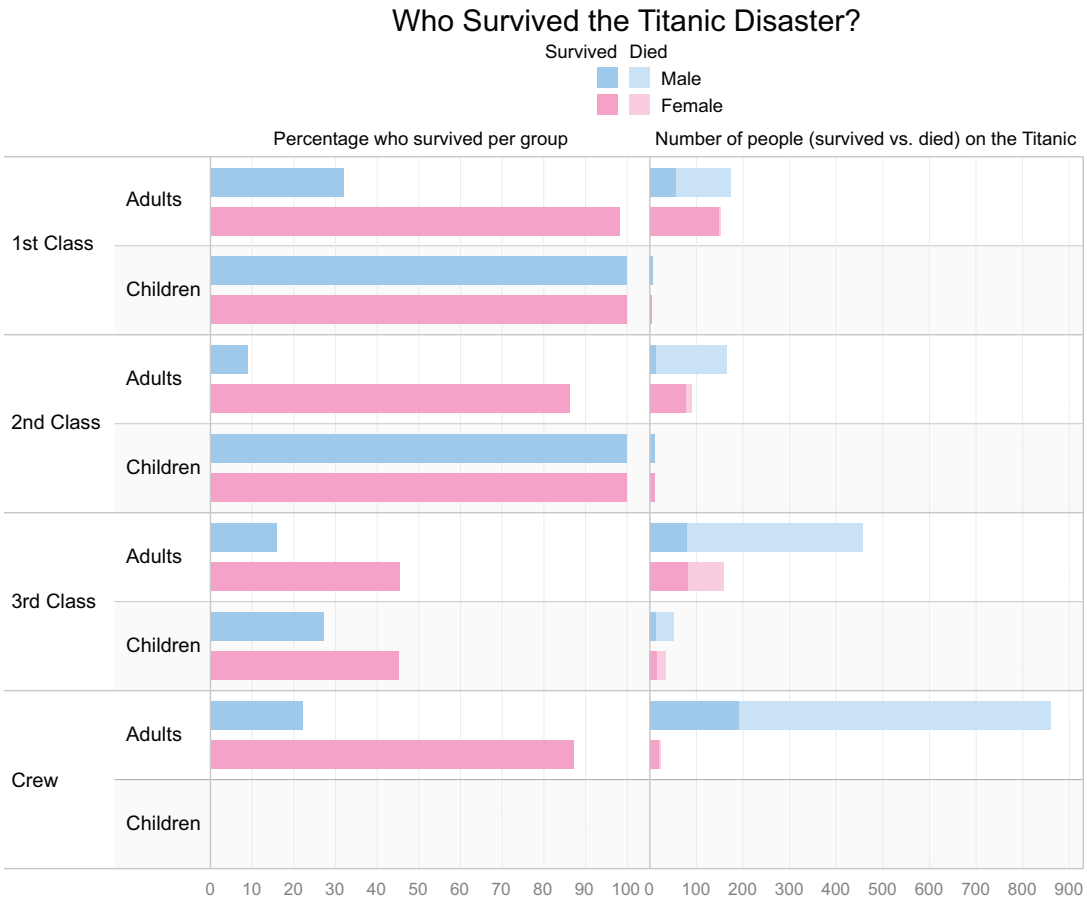
Now, here's the same information in bar graphs:



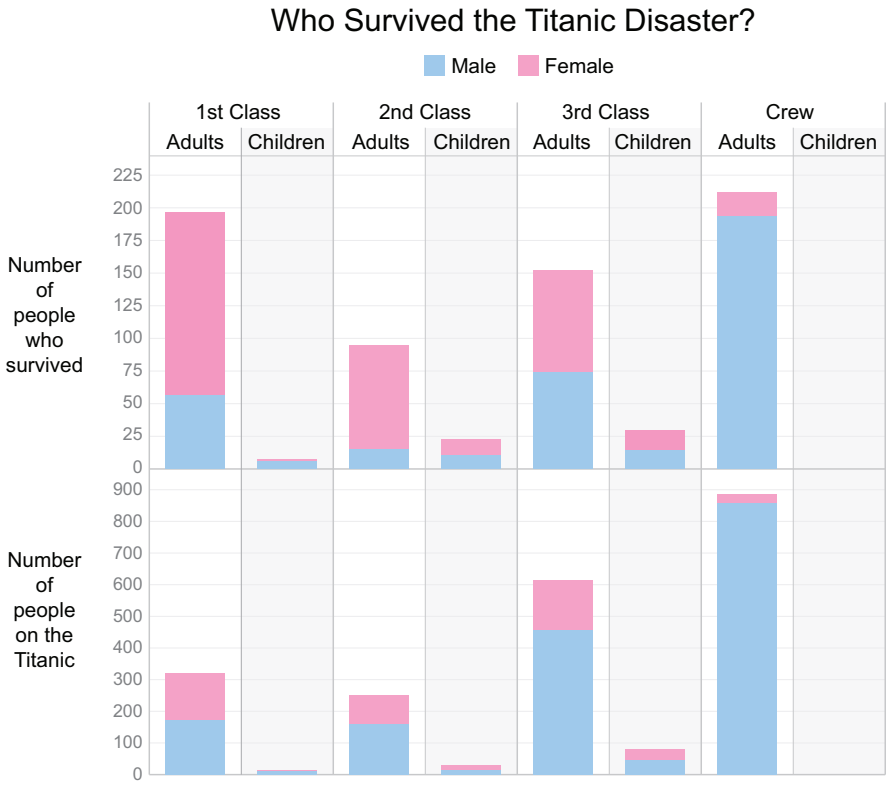
Now try to answer the previous questions using the bar graphs.

How many of the questions could you answer more easily using the mosaic plot versus the bar graphs? What questions of importance can you imagine asking that could be more easily answered using the mosaic plot?

I won't answer these questions for you, but I will share my answers with you. I've never found mosaic plots as useful as well-designed bar graphs for anything. I can, however, think of particular questions that could be answered more easily using bar graphs that are designed differently than the one above. For example, for some purposes, switching from vertical to horizontal bars (see following page) might work better.



Or, if we want to focus on the total number of people who survived of both sexes, a stacked bar graph like the one below would make this easier.



Mosaic plots aren't bad graphs, they just aren't optimal. With enough practice, you could use them reasonably well as a starting view of multidimensional part-to-whole data sets, but they aren't necessary, because everything that they reveal can be seen using bar graphs. Mosaic plots are the result of a well-intentioned effort to put complex, multidimensional, part-to-whole data into a single graph. While it's important to view all of the parts at once, it isn't necessary to squeeze them into a single graph. A coordinated set of bar graphs can reveal the same relationships in a way that can be more easily and accurately perceived and understood.

In addition to realizing that it isn't necessary to force everything into a single graph, it's also important to realize that no single view of data will ever answer every question. This is an underappreciated fact of visual analysis. Much time and effort is wasted trying to cram everything into a single view when moving fluidly from one view to the next usually works much better. There is greater value in a few graphs that our brains can easily perceive and interpret, empowered with the ability to quickly design them in various ways, than there is in a large assortment of graphs that includes many that are perceptually hobbled. Rather than adding every graph imaginable to BI products, I encourage software vendors to stick with the graphs that work best and empower them with an interface that allows those graphs to be easily designed in various ways, at the speed of thought. Providing graphs that are rarely useful complicates the interface with unnecessary choices and encourages people to use those graphs in ineffective ways. Keep it simple. More is often worse in the world of visual analytics.

Discuss this Article

Share your thoughts about this article by visiting the Are Mosaic Plots Worthwhile? thread in our discussion forum.

About the Author

Stephen Few has worked for nearly 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 three books: *Show Me the Numbers: Designing Tables and Graphs to Enlighten*, Second Edition, *Information Dashboard Design: The Effective Visual Communication of Data*, and *Now You See It: Simple Visualization Techniques for Quantitative Analysis*. 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.