Paper 361-2013

Make a Good Graph

Sanjay Matange, SAS Institute Inc.

ABSTRACT

A graph is considered effective if the information contained in it can be decoded quickly, accurately and without distractions. Rules for effective graphics – developed by industry thought leaders such as Tufte, Cleveland and Robbins – include maximizing data ink, removing chart junk, reducing noise and clutter, and simplifying the graph.

This presentation covers these principles and goes beyond the basics, discussing other features that make a good graph: the use of proximity for magnitude comparisons, nonlinear or broken axes, small multiples, and reduction of eye movement for easier decoding of the data. We also examine ways in which information can be obscured or misrepresented in a graph.

INTRODUCTION

We live in a world with ever increasing amount of data created from various sources and every transaction we incur. Health statistics, financial data, and credit card transactions are piling up in data warehouses. We want to analyze this data and form conclusions that can guide our actions. However, interpreting the results of our analyses from raw numbers is inefficient.

On average, a person is able to work with or retain only a few chunks of information at a time. On the other hand, the visual cortex shown in Figure 1, the largest part of the brain devoted to one function has evolved into a high-speed parallel system to rapidly interpret huge amount of visual information. Converting raw numbers into a visual representation enables us to use the brain's high-speed system for visual perception, resulting in faster processing of the data.

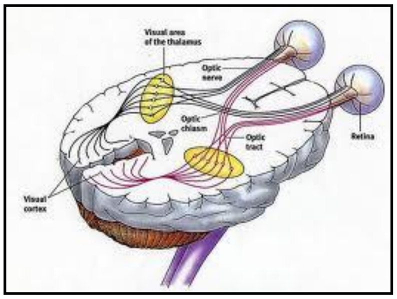


Figure 1 – Human perception

Even just a few years ago, investors would sit in the local office of a brokerage firm, watching the stock ticker. Latest trades of various stocks would stream by, and the experienced investor would be able to discern a trend in the data. This was a skill acquired over years of training and the casual viewer would see nothing but a jumble of numbers.

Normally, there are multiple stocks being actively traded at one time, so a jumble of numbers are scrolling by at one time, usually labeled and color coded based on the up or down tick. At the top of Figure 2, we have color-coded the stock name for clarity.

Often, what is most important to the investor is not the actual price of a stock, but whether it is trending up or down. To help the investor glean this information, we can represent the data visually, as a scatter plot with stock as the group as shown in the 2nd graph. Now we can start seeing some information in the data.

Because this data is a series over time, we can plot the data as a series plot grouped by stock as shown in the 3rd graph from the top in Figure 2. The stock price history is clearly visible now.

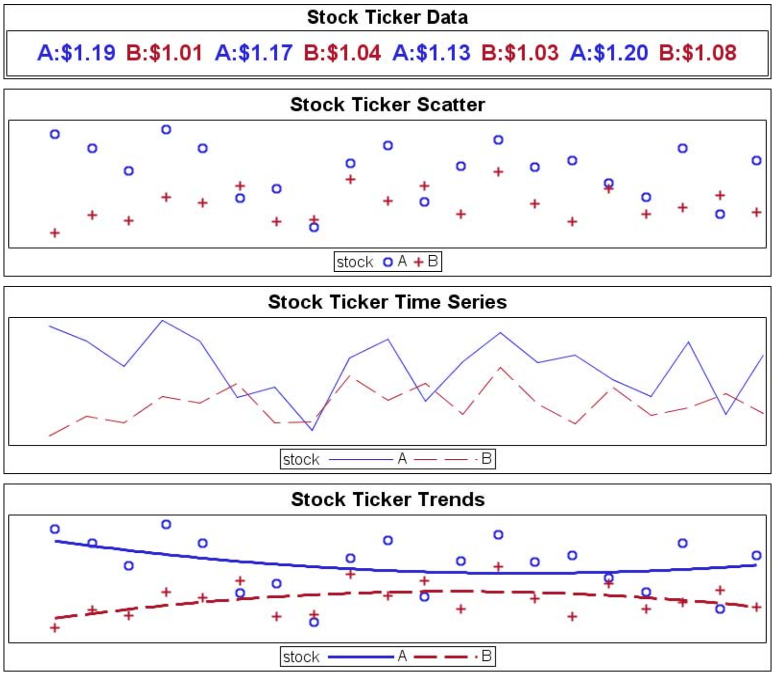


Figure 2

Lastly, we replaced the series plots with fit plots of degree=2. Alternatively, you can add moving averages. Now the investor has useful knowledge on which decisions and actions can be based. In the example above we have gone from raw data to information to knowledge by using graphs to represent the data in a form that is easier to absorb. Figure 3 shows the SAS® 9.2 SGPLOT code used to create some of the graphs in Figure 2.

But, what makes for a good graph? How do we ensure the information in the graph is communicated easily, quickly and accurately to the consumer? In this paper we will dive deeper into this topic.

Thought leaders in the field of data visualization have formulated some principles for creating effective graphs based on the science of visual perception and good practice. We will cover the topic in three parts as follows:

- visual perception science
- good practices for creating graphs
- Instructive use cases

VISUAL PERCEPTION

Figure 4 shows a schematic cross section of the eye. Note the location of the fovea, directly in the line of vision from the pupil to the retina.

The optic nerve carries the signals from all the photoreceptor cells to the visual cortex of the brain. At the center of the optic nerve is the blind spot, where there are no photoreceptor cells, so any part of the image that falls here cannot be perceived.

The retina of the eye contains two types of cells, the rods and the cones. The rods are sensitive to light intensity but not color. The cones are sensitive to the color of the light, and have a lower sensitivity. The cones have a higher resolution and can perceive fine detail in the image.

The fovea has a high concentration of cones and hence is the region where the image can be seen with higher detail. The fovea is responsible for the sharp central vision necessary for attentive activity such as reading, driving, and so on. The parafovea and the perifovea are farther away from the center and have a reduced number of cones, and thus have a lower resolution. About 50% of the nerves carry information from the fovea to the visual cortex.

How can we use the science of visual perception to make a good graph?

GRAPH FEATURES

While a graph has many uses, two of the common features of a graph that we will discuss here are as follows:

- group classification
- magnitude comparisons

Group classifications: In his book *Information Visualization: Perception for Design*, Colin Ware discusses the concepts of attentive and pre-attentive vision.

Attentive vision requires us to move the eye over the scene to carefully examine it. Figure 5 shows two visuals that look qualitatively similar but are not. The illustration on the left has red squares and blue circles. The illustration right also

```
/*--Display of Stock Values--*/
title 'Stock Ticker Data';
proc sgplot data=ticker(obs=8)
     noautolegend;
  scatter x=j y=k / markerchar=label
    group=stock;
  yaxis display=none;
  xaxis display=none;
  run;

/*--Display Values as Time Series--*/
title 'Stock Ticker Time Series';
proc sgplot data=ticker;
  series x=i y=value  / group=stock;
  yaxis display=none;
  xaxis display=none;
  run;

/*--Display Values with Fit Plot--*/
title 'Stock Ticker Trends';
proc sgplot data=ticker;
  reg x=i y=value / degree=2
      group=stock;
  yaxis display=none;
  xaxis display=none;
  run;
```

Figure 3

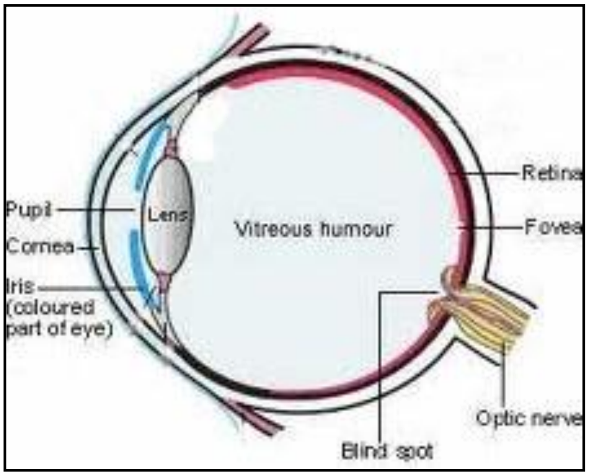


Figure 4

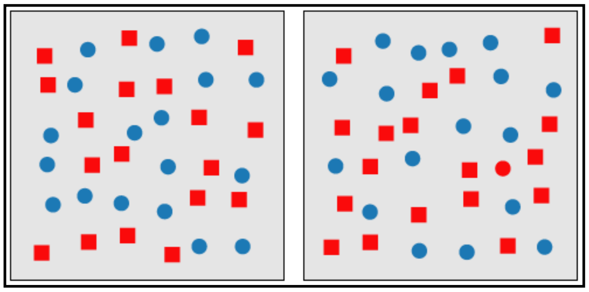


Figure 5

has red squares and blue circles, but it also contains a red circle. Do you see it? If you scan the illustration on the right carefully, you will see it.

Pre-attentive vision, on the other hand, occurs almost instantaneous.

Typically, any eye movement takes about 200 millisecond. If the task of decoding the information in a graph takes under 250 millisecond. that action is considered pre-attentive. Figure 6 shows an example of what can be decoded with pre-attentive vision. The key information in the visual can be decoded very quickly, without the need to carefully scan the visual.

The usage of pre-attentive features can greatly speed up the decoding of information in a graph. The following features used for group classifications can be detected by pre-attentive vision:

- discrete color
- discrete marker shape
- line pattern
- marker orientation.

Magnitude comparisons: In his book *Sensation and Perception*, Bruce Goldstein presents the results of studies that measure the perceived response to a stimulus. Figure 7 shows the perceived response for different levels of stimulus for three different cases.

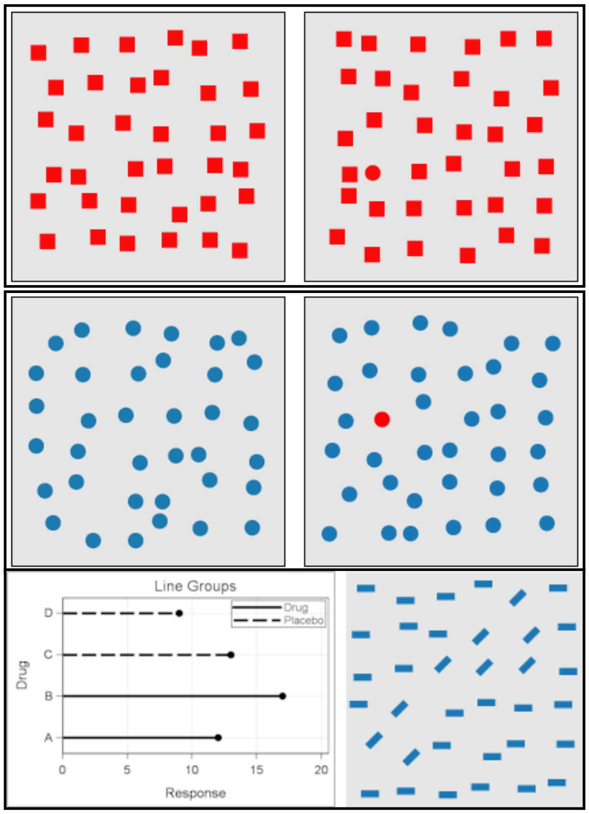


Figure 6

The graph plots the intensity of stimulus on the X axis, and the perceived response on the Y axis. As we can see from the plot for the line length, the perceived response is almost linear with a slope close to 1.0.

Steven's Power Law shows the relationship between perceived response (P) and stimulus (S) as follows:

$$\mathbf{P = K * S^N}$$

For linear line lengths, K and N are close to 1.0 indicating that linear distance is an excellent representation for magnitude such as in a dot plot. This is especially true when these are drawn from a common baseline as in a bar chart or needle plot.

It is important to note that brightness is a poor representation of magnitude, as the perceived response for increasing brightness falls off exponentially. Thus, usage of color brightness is not recommended as a good quantitative representation of magnitude.

Figure 8 shows some recommended representations of magnitude. In general, the following are recommended:

- linear distance from a common baseline
- linear distance
- marker size

When magnitude is represented as area or volume, we tend to underestimate the response as follows.

- Representing magnitude as area has K=0.7.
- Representing magnitude as volume has K=0.6.

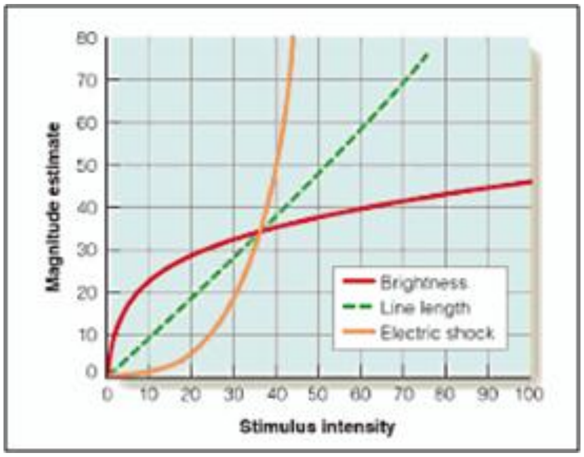


Figure 7

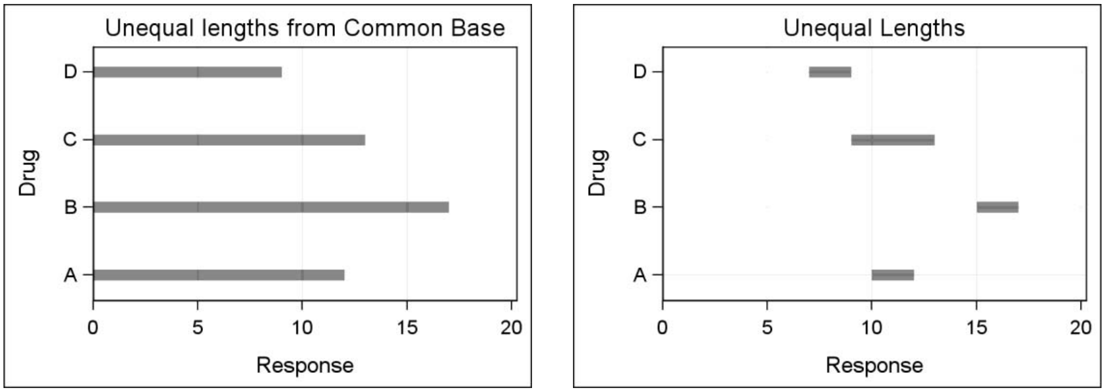


Figure 8 – Representation of Magnitude as Linear Length

Area and color for representation of magnitude. Often, we see use cases where magnitude is represented by area as in a Tree Map or Tile Chart as shown in Figure 9. In this case, this is useful as each rectangle represents the sum of magnitude of the hierarchy below it. This visual is able to represent both magnitude and relation between parent and children nodes.

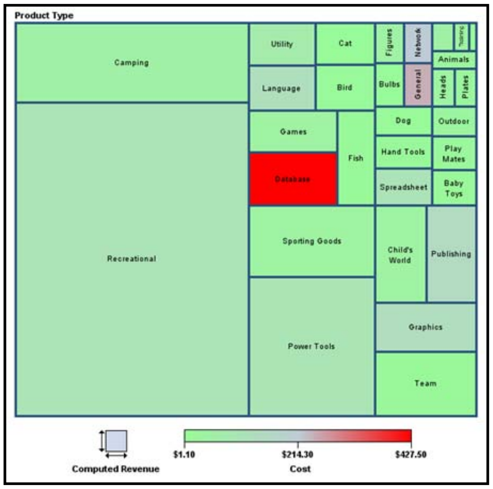


Figure 9 – Tile Chart

In addition, another response can be represented by the color of the rectangle. The graph in Figure 9 displays the revenues for each product hierarchy as area of the rectangle. The cost for each product is displayed by the color of the rectangle.

In this case, the use of color provides a qualitative representation of the magnitude, and not quantitative. The legend at the bottom of the graph provides a range of values represented by the colors. However, there is really no way to make an accurate estimate of the magnitude based on just the color.

Use of accessible colors. When using colors for either group classification or representation of magnitude, it is important to consider the accessibility of colors.

It is estimated that between 7 and 10% of the US population has some deficiency in perception of color. The most common type of this deficiency is lack of discrimination between red and green colors. This affects about 5% of the male population. Figure 10 shows a pattern of red and green dots with a number embedded in the middle, a common test to evaluate this type of deficiency. .

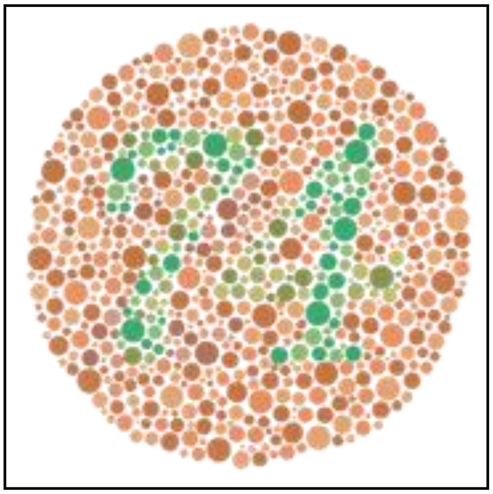


Figure 10 – Ishihara Test

Cynthia Brewer's website http://colorbrewer2.org is a good resource addressing the topic of accessible colors.

Recommendations for group classifications:

- Use discrete color only.
- Use discrete marker shapes only.
- Use line patterns and marker orientation.

Recommendations for representation of magnitude:

- Use linear distance from common baseline.
- Use linear lengths.
- Use marker size.

GOOD PRACTICES FOR MAKING A GRAPH

Based on the science of visual perception and their experience over the years regarding what works and what does not, thought leaders in the field of analytical data visualization (such as Edward Tufte, William Cleveland and Naomi Robbins) have presented a set of recommendations for creating effective graphs. In her book *Creating More Effective Graphs*, Naomi Robbins describes how to evaluate the effectiveness of a graph as follows: “A graph is more effective than another if its quantitative information can be decoded more quickly or easily by most observers”.

Edward Tufte’s recommends, and I paraphrase, that we simplify, increase data ink, and avoid chart junk.

Let us apply some of the science of visual perception, and the recommendations of the authors mentioned above to the task of producing a good graph.

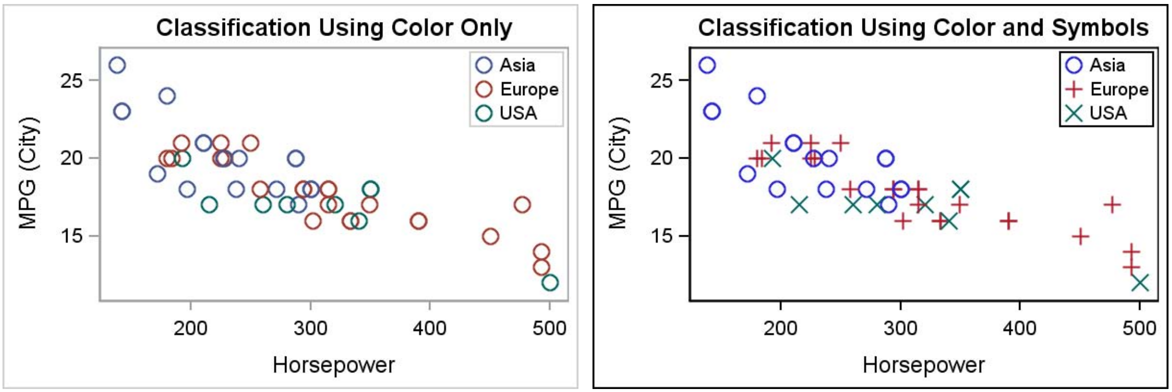
Here are some features of a good graph.

1. Use pre-attentive features for group classifications.
2. Use linear distance for the representation of magnitude.
3. Increase the proportion of data ink.
   a. Reduce clutter
   b. Reduce noise.
   c. Avoid chart junk.
4. Reduce the amount of eye movement required to decode a graph
   a. Increase proximity for comparisons.
   b. Use direct labeling where possible instead of legends.
5. Chunking, micro maps.
6. Small multiples.
7. Display the data directly.

1. Use pre-attentive features for group classification.

The graph on the left in Figure 11 shows the usage of color only for classification. The three group values are represented by three distinct colors but the same marker symbol. This usage of color only works well when using a color medium and can be easily achieved by using the new HTMLBlue style available with SAS 9.2.

In the graph on the right, each group value is represented by a unique color and a unique marker symbol. This is useful for a color print, but is necessary when using a monochrome output. This can be created by using the default LISTING style.



```
ods listing style=htmlblue;
```

```
ods listing style=listing;
```

Figure 11- Use pre-attentive features for group classification

2. Use linear distance for representation of magnitude.

Figure 12 shows three different ways to represent magnitudes in a graph using a pie chart, a bar chart, and a dot plot. All graphs show the sum of the actual revenues by product category. The goal is to see which of these visuals is effective for representation of the magnitude.

For the pie chart, the categories are labeled in each slice. The label for the response value is not shown. For a fair comparison, the response axis for the bar chart and the dot plot are suppressed.

Pie chart: In this graph, the response values are represented as the angle subtended at the center of the pie for each slice. Looking at the slices of the Pie Chart, it is not easy for the average reader to determine which category has the highest response value. This is because it is not easy to compare magnitudes as angular distances, especially from different baselines.

Also, as the number of slices increase, it becomes increasingly difficult to label the slices inside the pies. Often in such cases the slice labels are moved outside the pie, thus reducing the size of the pie itself.

The Pie Chart is not highly effective in such use cases where the comparison is between categories. However, it can be a very effective representation in cases where a part-to-whole comparison is desired. This is discussed later in this paper.

Bar chart: In this graph, the response values are represented as a linear length from zero baseline along the response axis (vertical axis in this case) for each category shown along the horizontal axis. Normally, the response axis is labeled with a linear scaled numeric axis. Grid lines can be provided for each major tick value on the response axis.

Even without any tick values on the response axis, it is easy for the average user to determine which category has the highest response value. It is also possible to make a reasonably accurate estimate of the relative magnitudes. This is borne out by the research we discussed earlier showing the perceived response for a linear line for a given stimulus is very effective.

Dot plot: In this graph, the response values are represented as a linear distance along a horizontal response axis for each category along the vertical axis starting from the top. Normally, the response axis is labeled as a linear scaled numeric axis. Grid lines can be provided for each major tick value on the response axis.

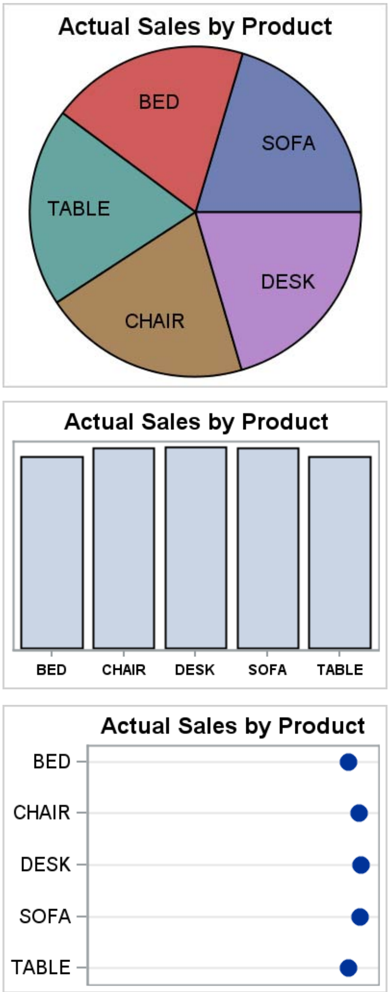


Figure 12

For this comparison we have set the baseline value to zero as shown in Figure 12. By default, the response axis does not have a baseline of zero, and the response axis extends only to cover the range from the minimum to the maximum response value. This feature of the dot plot makes it a better fit when a non-linear scaled response axis is used. This is discussed later in this paper.

Even without any tick values on the response axis, it is easy for the average user to determine which category has the highest response value. It is also possible to make a reasonably accurate estimate of the relative magnitudes. This is borne out by the research we discussed earlier showing the perceived response for a linear distance for a given stimulus is very effective.

3. Increase the proportion of data ink.

Often we see graphs where the key information is obscured by unnecessary clutter, noise, and other artifacts. They create distractions in the graph and draw the eye away from the essential information.

a. **Reduce clutter:** Figure 13 shows graphs of mileage by horsepower. The graph on the left includes a large number of tick values on the axes, along with strong grid lines on both axes. This distracts from the information in the graph. This graph sets the axis origin at (0. 0), that takes away valuable space in the graph.

The graph on the right shows the same information with a smaller number of tick values on both axes. The axes have linear scaling anyway, so it is not really necessary to have a large number of tick values on each axis. As a consequence, the number of grid lines is also reduced, resulting in a clean look. The axis range is enough to cover the data range, thus providing more of the valuable space to the data.

The graph on the right represents the default graph output from the SG Procedures and the Graph Template Language (GTL). You do not have to do anything extra to get the good, uncluttered graph.

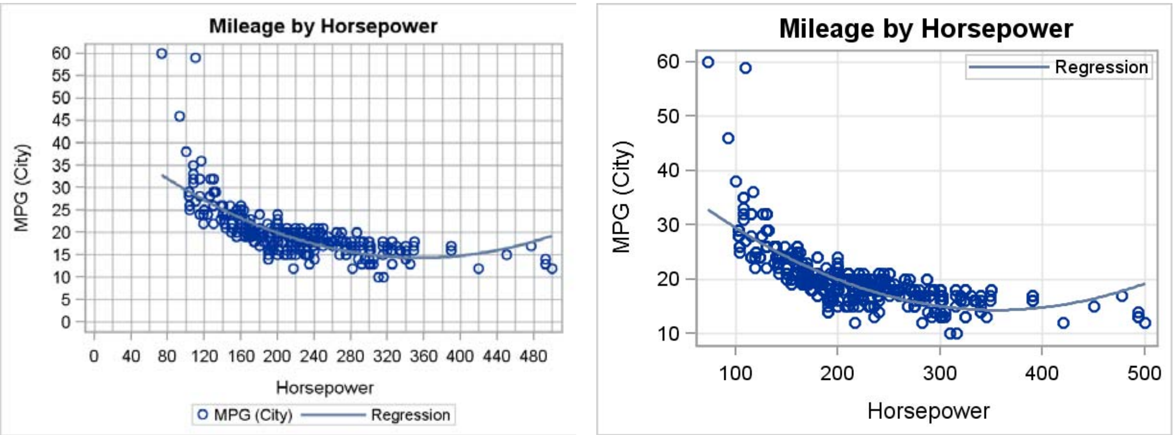


Figure 13

b. **Reduce noise:** Figure 14 shows bar charts of oil production by year at a production facility. Oil, gas, and water are plotted in the graph as stacked bars. The graph on the left includes an image of an oil production rig behind the data in the graph and the bars themselves have a 3-D projection.

The graph on the right shows the same information using a 2-D stacked bar chart. The background image is removed, allowing the data to be clearly visible without any distractions.

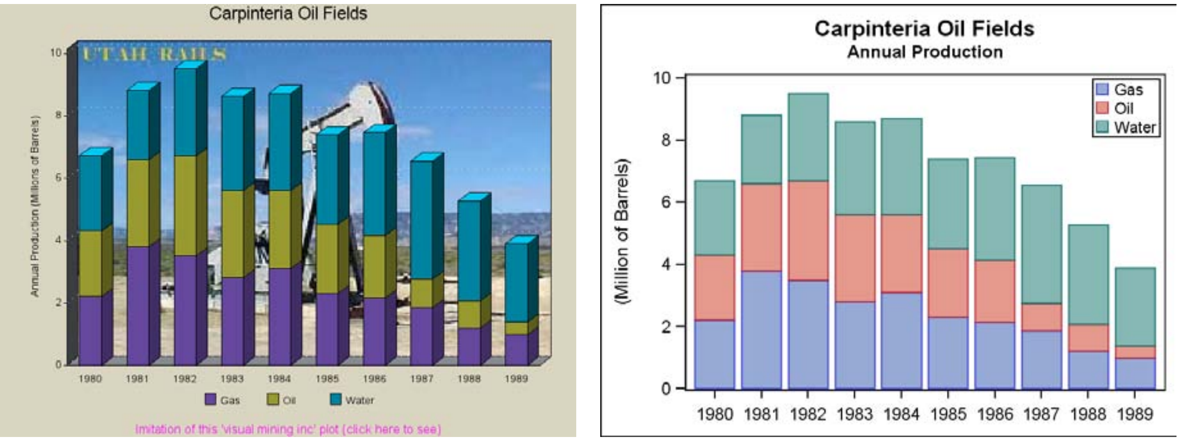


Figure 14

c. **Avoid chart junk:** The term "chart junk" was coined by Edward Tufte in his 1983 book *The Visual Display of Quantitative Information.* Tufte wrote:

"The interior decoration of graphics generates a lot of ink that does not tell the viewer anything new. The purpose of decoration varies — to make the graphic appear more scientific and precise, to enliven the display, to give the designer an opportunity to exercise artistic skills. Regardless of its cause, it is all non-data-ink or redundant data-ink, and it is often chart junk."

By this definition, the interior image behind the data and 3-D shape of the bars in the graph on the left in Figure 14 also qualifies as chart junk.

Figure 15 shows three graphs showing the mean mileage by vehicle type. The graph on the top uses an extruded 3-D-ooking bar shape to represent the data. This visual is often used in marketing brochures, and is often embellished with textures, lighting and shadows.

The problem with this visual is it reduces the effectiveness of this graph and makes it harder to accurately decode the data in the graph. For example, in this graph, it is harder to estimate the magnitude of the tallest bar. It is also hard to compare the values for the SUV and Truck.

On the other hand, the simple 2-D graph in the middle is very effective for representation of the magnitudes and for comparisons across the categories. From this graph, the average viewer can see that mean mileage for Trucks is just a little higher than SUV.

Note: With SAS ® 9.3, Statistical Graphics (SG) Procedures and GTL provide an option called DATASKIN to render nicer looking 2D bars as shown later in Figure 23. Such effects are off by default. It is up to the user to use these if appropriate.

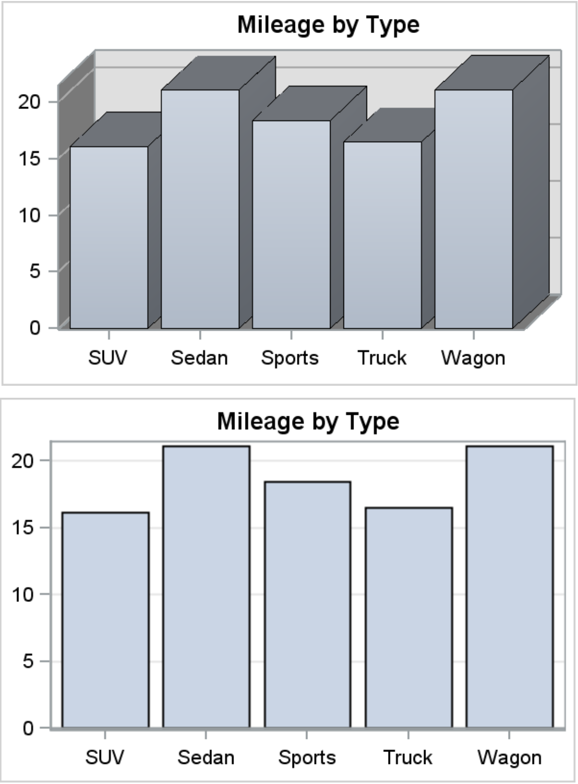


Figure 15

4. Reduce eye movement

As we learned in the section on visual perception, attentive decoding of a visual requires the viewer to move the eyes so that the feature to be examined falls on the fovea, the region of the eye with the highest resolution.

a. **Increase proximity for comparisons**. A graph is easier to decode if it requires less eye movement to examine the features. In the graphs in Figure 15, it would be easier to compare the magnitude for SUV and Truck if these categories are placed closer in the graph.

In Figure 16, the graph shows class strengths by college with class as the group variable. This is created by a user to compare the number of freshmen in different colleges. As we can see, the bars for freshmen for each college are far apart, and this arrangement is not ideal for such comparisons.

The graph in Figure 17 displays the class strength by class with college as the group variable. In this graph, all the bars for freshmen from different colleges are placed close together for easier comparison. Comparison of the strengths of other categories such as Graduate, Transition, and Transfer are also easy to compare side by side.

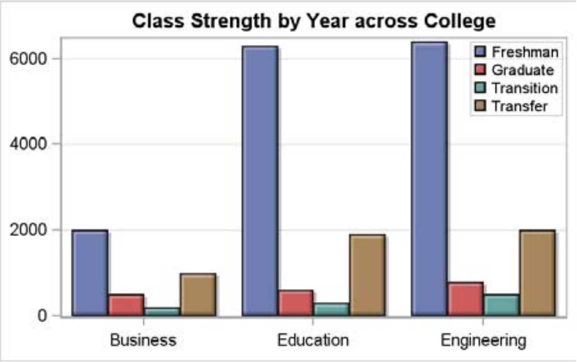


Figure 16

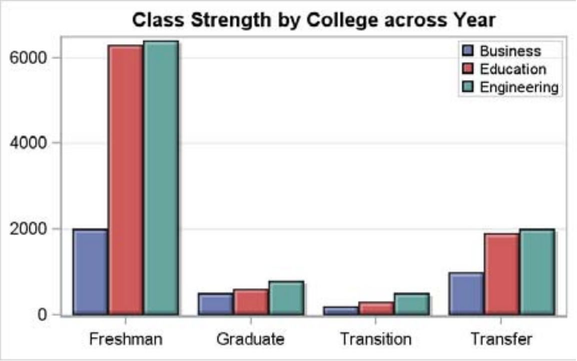


Figure 17

b. **Use direct labeling.** Just as magnitude comparisons are easier when the values are closer together, similarly placing identifying labels close to the object itself is beneficial. This reduces the eye movement required to decode the graph and thus makes it easier.

Figure 18 shows two series plots of the revenues by month for five different products. Traditionally, a legend is placed at one side of the graph, usually at the bottom to identify each series as shown in the graph on the left in Figure 18. In this arrangement, comparing the performance for Chair and Table requires us to move the eye to the legend to identify each curve and then decode the values. Similarly, comparing other values requires extensive eye movement.

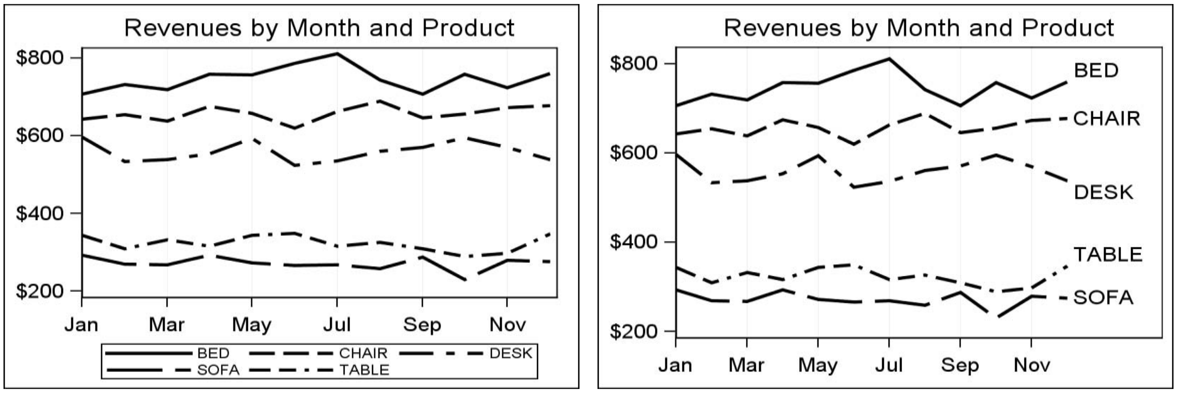


Figure 18

A better arrangement is shown on the right of Figure 18, where the series labels are placed adjacent to the curve itself. Now it is very easy to compare any two curves in the graph, and eye movement is significantly reduced. In this graph, we have effectively moved the legend right next to the objects themselves.

5. Chunking and micro maps

Chunking is a technique used by all of us to help in short term memory. Studies have shown that most people can remember a small number of items at a time, generally three to five.

So, if we are to absorb and recall more than 5 objects, it helps to chunk these into smaller groups. For example, a 10-digit phone number like 9195316753 is easier to recall when chunked into smaller pieces like (919) 531-6753.

A simple example is shown in Figure 19 to illustrate the point. A simple horizontal bar chart on the left side includes 15 car makes, sorted by descending mean mileage. The same graph on the right breaks up the graph in to smaller chunks, creating a less intimidating graph.

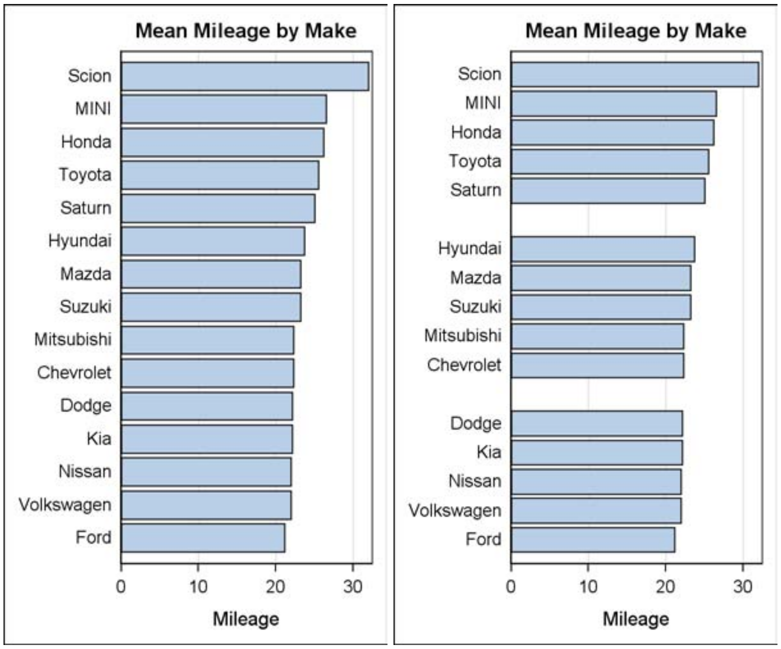


Figure 19

Micro maps: Another good example of chunking is the Micro Map array shown in Figure 20. In this graph we are viewing the rainfall in each region, and providing a visual key to where this region is in the map of USA.

One way would be to include one map of the US, identifying all the regions using nine different colors, and then including one horizontal bar chart of all the nine regions in descending amounts.

The graph in Figure 20 breaks up the graph into smaller chunks, thus making it easier to absorb the data. A small map is used to display three regions alongside the bar chart for those three regions.

This enables us to use only three colors to identify each region in the map and the bar chart, thus reducing the clutter.

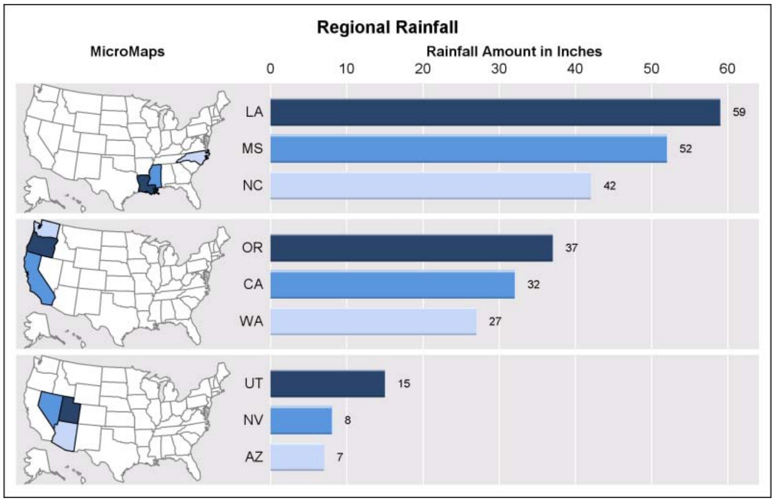


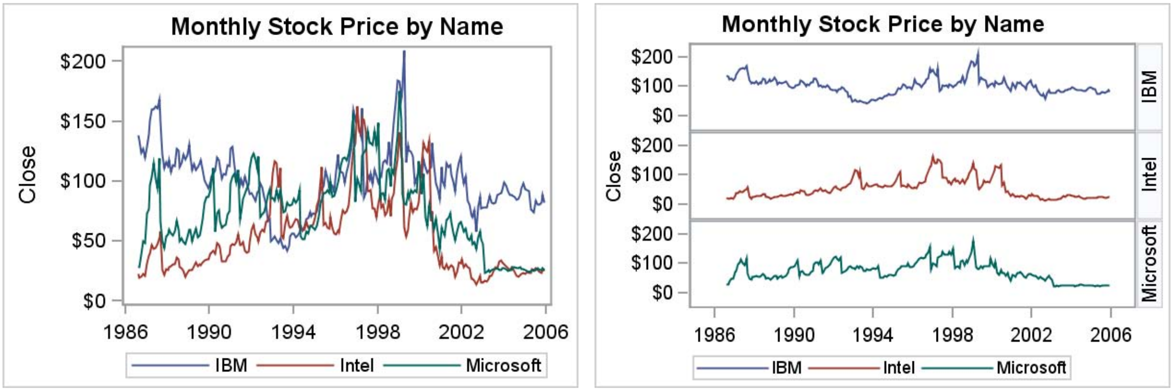
Figure 20

6. Small multiples

“Small Multiples” is a technique popularized by Edward Tufte to present information in small related chunks.

> “At the heart of quantitative reasoning is a single question: *Compared to what?* Small multiple designs answer directly by visually enforcing comparisons of changes and differences. For a wide range of problems in data presentation, small multiples are the best design solution.” – Edward Tufte (Wikipedia).

Figure 21 shows two graphs of the monthly closing stock prices over time for three different stocks from the SASHELP.STOCKS data set.



```
title 'Monthly Stock Price by Name';
proc sgplot data=sashelp.stocks;
  series x=date y=close /
    group=stock;
  keylegend / title='';
  xaxis display=(nolabel);
  run;
```

```
title 'Monthly Stock Price by Name';
proc sgpanel data=sashelp.stocks;
  panelby stock / layout=rowlattice
          novarname;
  series x=date y=close /
    group=stock;
  keylegend / title='';
  colaxis display=(nolabel);
  run;
```

Figure 21

The graph on the left uses a series plot by stock to display the three different curves in one graph. The SGPLOT procedure code is shown below it. The graph on the right shows the same data as a paneled graph using the stock name as the panel variable. The SGPANEL procedure code is shown below the graph.

The benefits of the graph on the right are clear, as the three curves are shown in separate cells on a common date axis. The graph has significantly less clutter and the information is easy to absorb.

7. Display data directly.

Often when collecting information, the data collected is of actions or processes that are directly measurable, such as hours worked, revenues, expenses, and so on.

When reporting on a derived value such as profitability of an enterprise, the measured values are revenues and expenses. So, when creating a graph of the measures, we often plot the values actually measured, leaving the derived value up to the user to determine.

This can sometimes lead to inaccuracies as shown in the graphs in Figure 22. In the graph at the top, we have plotted the revenues and expenses. The profit is the difference between those two values. So, what would be in your estimate the time of the steepest drop in the profits? It is not easy to estimate this from the upper graph.

This is, however, more clearly evident in the bottom graph, where we have plotted the actual profit in addition to the revenues and expenses. The light blue band represents the actual profit. Note how the profit takes a steep plunge at x=18.

This is not easy to discern in the upper graph because the eye estimates the difference between lines as the nearest distance, and not the vertical distance, which is the real measure of the profits.

Plotting the curve for profits directly is a better way to convey the information to your user.

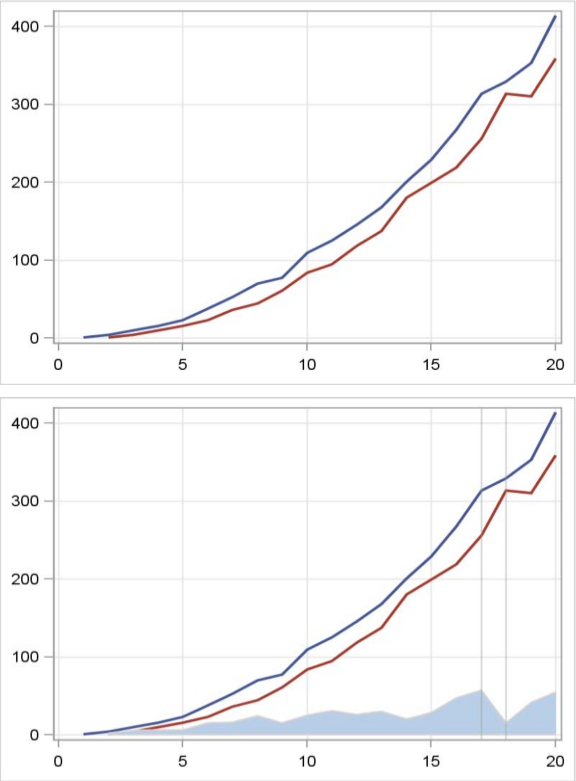


Figure 22

SOME INSTRUCTIVE USE CASES

Let us consider some real world use cases to illustrate the features of a "good graph".

1. Bar charts with nonzero baseline.

Figure 23 shows two bar charts for CPU performance for three different vendors. The graph on the top accentuates the difference in the magnitude by setting the baseline at 3.0. In this graph, "our" CPU appears to have a significant performance edge over those sold by other vendors. This is a standard marketing trick, and is clearly misleading.

The graph on the bottom shows the more accurate comparison between CPUs from different vendors

.

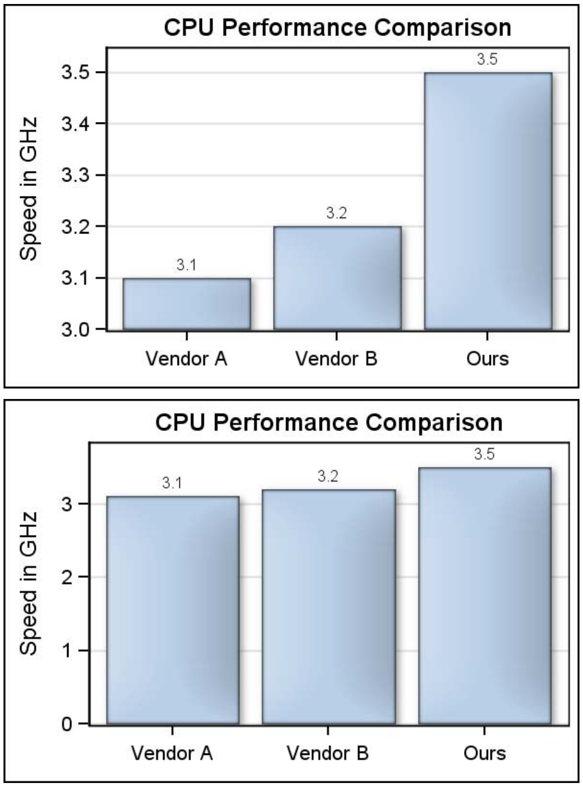


Figure 23

2. Bar charts with non-linear response axis

Often, in the Health and Life Sciences use cases, the measured responses for cell growth and other natural phenomenon have non-linear relationships and are best visualized using log scales with base 2**.** Figure 24 shows two bar charts showing breast cancer stages using bar charts. The one on the left is a traditional chart with a linear axis and zero baseline.

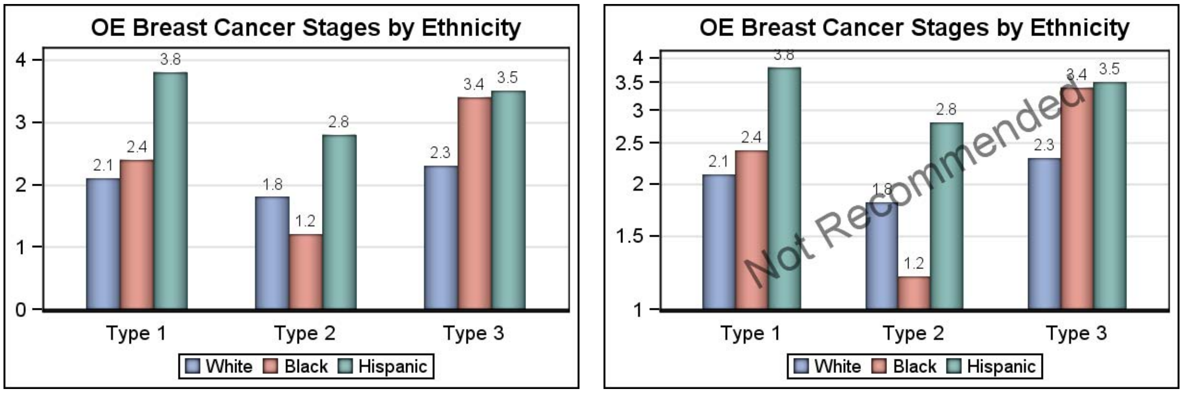


Figure 24

In this case, the responses have a non-linear relationship that the user wanted to represent as a log base 2 axis scales. For SGPLOT and GTL graphs, a log axis can be made only if all values on the axis are greater than zero. Therefore, I had to artificially set a baseline of 1.0, and axis type of log. This is rather arbitrary, and the shape of the bar charts looks quite different with different baseline values.

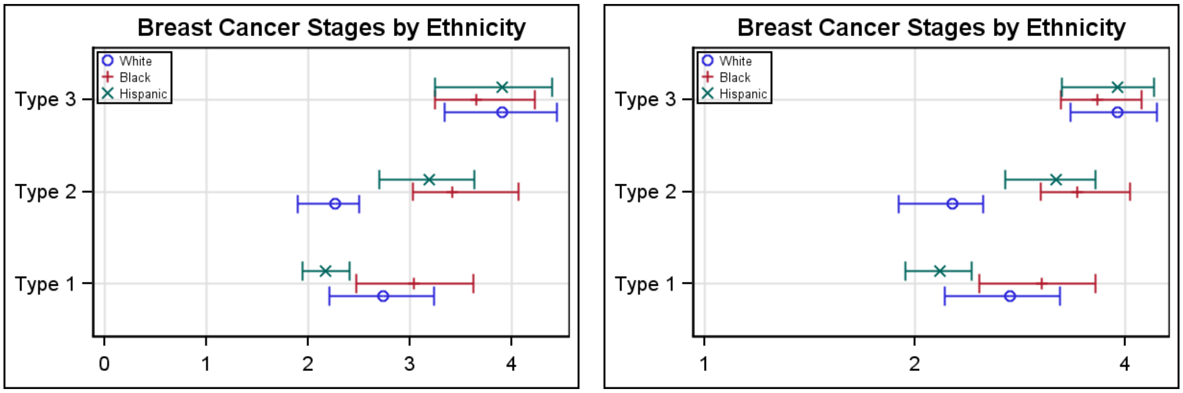


Figure 25

While non-linear axes are not recommended in general, exceptions are made in cases where the consumer is well aware of the use of such non-linear axes, and there is a significant benefit in the data representation. However, in this case, the recommendation by thought leaders in the field is to use a dot plot, where there is not a strong association with baseline as shown in Figure 25.

3. Comparing magnitudes with large differences

Another frequent issue is to compare magnitudes between categories having a large difference between the values. Figure 26 shows such a case where two of the categories (E and F) have response values that are an order of magnitude larger than the other four.

Visualizing data like this is a challenge, as can be seen in Figure 26. The presence of the large values forces the scaling of the graph to accommodate all the bars. This causes the smaller values to be

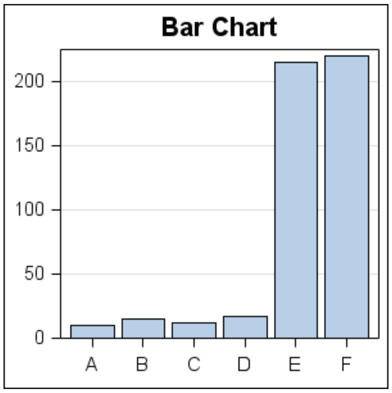


Figure 26

represented in a very small part of the graph, making it harder to compare among the smaller values.

One possibility is to use a log response axis, except we just got through discussing how that is not a good idea in the previous section. A common solution is using a “broken axis” where the response axis is broken into two (or more) parts, retaining only the data ranges that are of interest.

Figure 27 shows such a solution using the GTL lattice layout of two rows. We have populated the same bar chart in both the cells with a common external column axis.

Knowing the distribution of our data, we set the Y axis ranges for the lower cell from 0–25 and the upper cell to 200–220, creating the appearance of a broken the axis. We set the tick values for both the Y axis with equal tick intervals.

This graph enables the shorter bars to grow in size, so they can be more easily compared with each other. The taller bars are visible in the upper cell for easier comparisons with each other.

One problem with this graph is that while it provides us the ability to focus in on the shorter and taller bars, there is a loss of context where we lose the sense of comparison between the two sets. Without carefully examining the broken Y axis, one could lose sight of the fact that E and F have values much, much larger than the other bars.

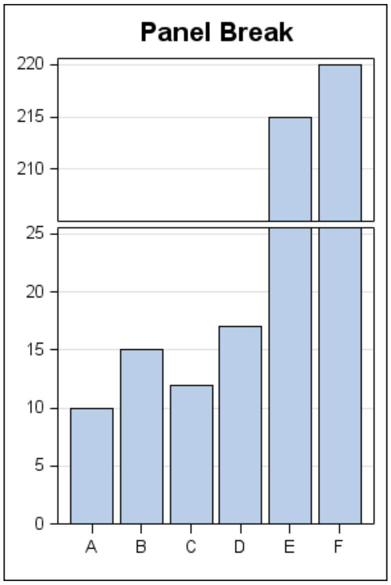


Figure 25

Retaining context and focus: Another creative solution to this problem is suggested in Figure 28. The goal here is to retain both context and focus. We want to be cognizant of the large disparity in the magnitudes, while also have the ability to compare between the shorter bars.

In this solution we have again used the GTL lattice layout, and populated the same graph in the upper and lower cells, with a common external column axis as in Figure 27. In this case, the upper cell shows the entire extent of the Y axis. For the lower call, we have set the Y axis range as 0–25.

In addition to this, we have used the high-low plot to display the bars in the lower cell. The high-low plot enables us to display bar caps for those bars that have been clipped by the axis range. This is a special behavior of the high-low cap with SAS 9.4 and does not need any special coding. The same effect can also be achieved using SAS 9.3 by setting values for these bars explicitly with the HIGHCAP option.

In my opinion, the solution in Figure 28 provides the ability to focus on the shorter bars, while retaining overall context of the visual.

See my blog post in “Graphically Speaking” on “Broken Y_Axis” posted in June 2012 for full SAS code.

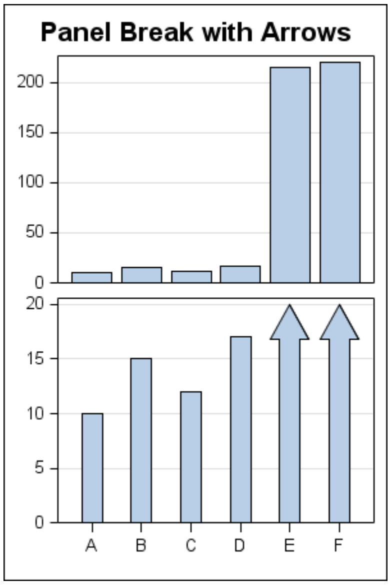


Figure 26

4. Good pie charts

Any discussion on usage of pie charts for data presentation can get contentious, with strong opinions expressed. Many thought leaders have suggested that pie charts are not suitable for effective presentation of data. In her book *Creating More Effective Graphs*, Naomi Robbins echoed opinions by Edward Tufte and William Cleveland, who do not favor the use of pie charts, by saying, “The only thing worse than a pie chart is several pie charts.” Naomi Robbins includes 3-D pie charts in the same category.

When used to display multiple category (slice) values, pie charts have difficulties. which include the following:

- It is difficult to compare magnitudes of one slice to another when represented as angles.
- This gets harder as each slice starts from a different base angle.
- Labeling pie slices is hard, and gets harder with increasing number of slices.

Part-to-whole relationship:

Pie charts are popular, and do seem to work well for displaying part-to-whole relationship, as shown in Figure 29.

Figure 29 shows two cases where the share of Sedans or SUVs to the entire population can be intuitively displayed using a pie chart.

This visual appears to be more satisfying when the center line axis of the two slices is either vertical or horizontal. One could attribute this to the principle of Gestalt perception based on the law of *pragnanz*, which says we prefer symmetric and simple organizations (Wikipedia).

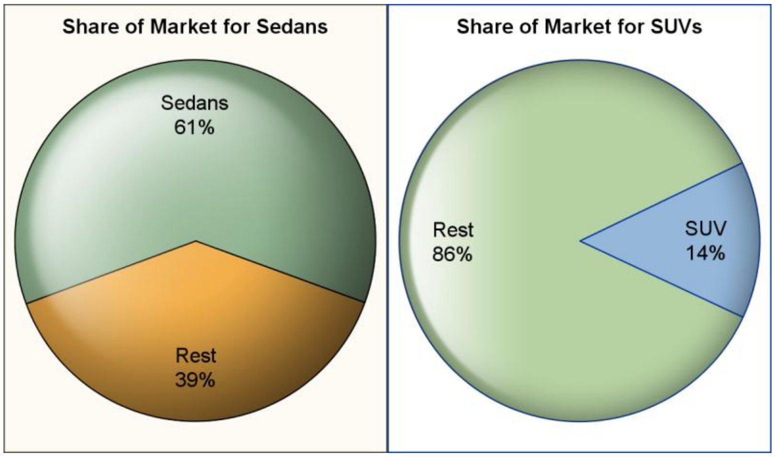


Figure 27

5. Good graph sizing

Starting with SAS 9.2, graphs can be automatically obtained from many SAS analytical procedures. Also, you can use the new Statistical Graphics procedures, Graph Template language and the ODs Graphics Designer application to create custom graphs.

The default size for these graphs is usually 640x480 pixels. Often, you might need to include a graph created from these procedures or applications into a report or paper, as shown in Figure 30.

Embedding the default size graph into a 3inch wide space results in the original graph being scaled down to fit the space. All elements in the graph are linearly scaled down, often resulting in a graph where the text is difficult to read. This is the case for the graph on top in Figure 30.

When you know the ultimate destination and size of the graph, it is better to render the graph to the final size and dpi required. The lower graph in Figure 30 was rendered to a size of 3.25 inch with 200 dpi using the options on the ODS GRAPHICS and ODS destination statements. When the graph is rendered at the correct size, the font and marker sizes are more readable in the small space.

Suggestion: Often, you need to embed a graph into a document of paper. For the best result, render the graph to the final image size with high dpi.

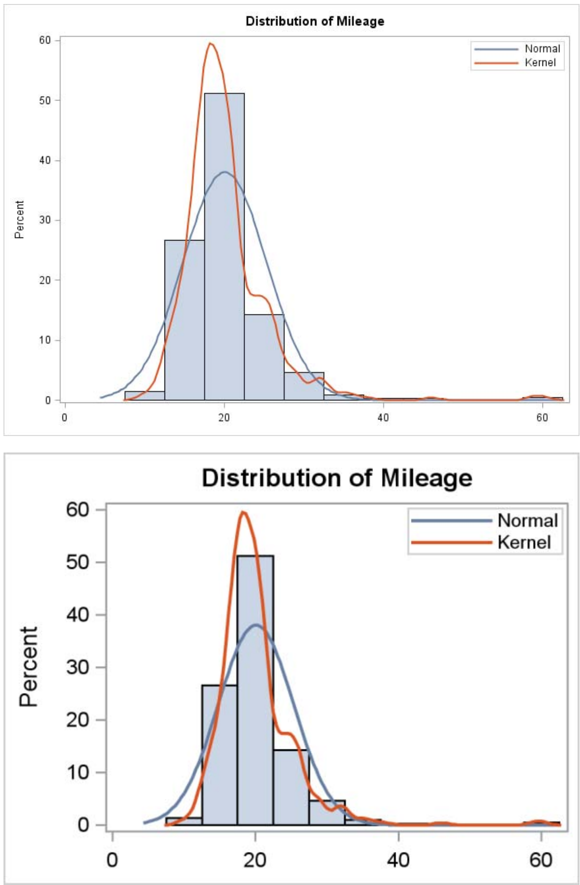


Figure 28

6. Improve graph readability

The graphs created by SAS analytical procedures, SG Procedures, GTL, and ODS Graphics Designer use principles of effective graphics by default to create good graphs. However, often small changes can be made to a graph to improve its readability. Figure 31 shows a sequence of graphs showing vehicle mileage by type and origin. Origin is used as the group role.

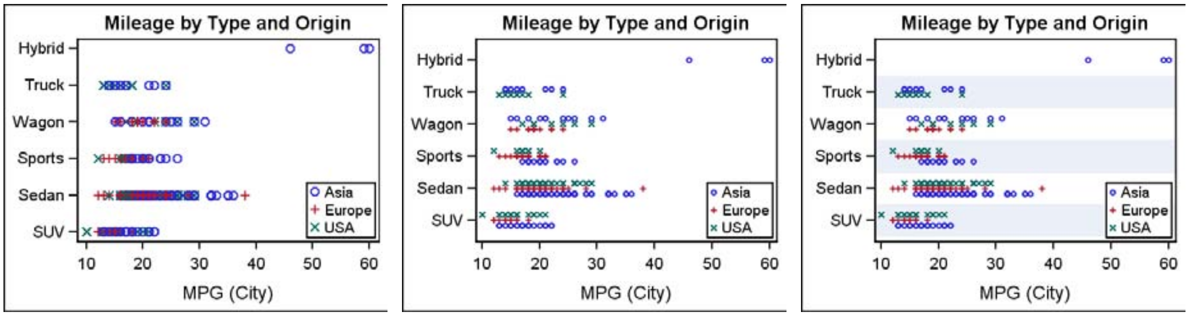


Figure 29

The first graph on the left shows the default rendering, where the different groups are drawn coincident on the category midpoint. Markers for all groups for any type (say, SUV) are drawn on the same Y axis midpoint location and therefore are jumbled together.

The second graph uses the GROUPDISPLAY=CLUSTER option available with SAS 9.3. With SAS 9.3, groups for all plot types can be drawn overlaid or clustered. This improves the graph, and the observations are more clearly visible.

The third graph uses alternate horizontal bands to delineate the clusters for the categories. This is helpful to guide the eye across the graph, and also helps to “chunk” the graph thus making it more readable.

The alternate category bands are a new feature that will be available with SAS 9.4. However, while this makes it very easy and convenient, you can still create alternating bands with SAS 9.2 or SAS 9.3 by using techniques described in the Graphically Speaking blog article “Forest Plot with Subgroups.”

This same technique is used to improve upon the readability of the forest plot shown in on top in Figure 30. In this graph, the data is sub grouped by various categories. The graph on top is a common version. As you can see, it is difficult to align and absorb all the data.

Alternating bands, one for each sub group, are used to help align the data. These graphs are quite complex, and it is not possible to render them to a three-inch space. So, here I am forced to include “scaled” versions for illustration purposes only.

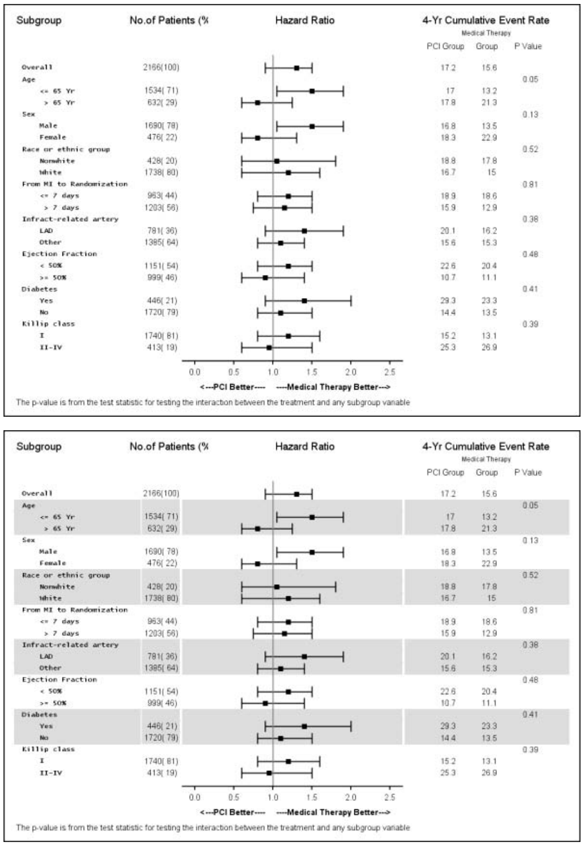


Figure 30

CONCLUSIONS

Graphs are increasingly seen as integral part of an analysis or a study. Graphs enable us to convert raw data into information, and information into knowledge that can be used for decisions and actions. It is important that the graphs we create are effective in communicating information.

A graph is effective if the information in the graph can be easily and accurately decoded by the user. To help in this task, thought leaders in the field of data visualization have formulated rules that can help us create a good graph. These rules are based on the science of visual perception and on studies on human perception. We can use these rules and guidelines to create graphs that convey the information accurately with minimal room for misinterpretation.

You can download a copy of the code and data used in this paper from the SAS Technical Papers and Presentations site at http://support.sas.com/rnd/papers/index.html. Find the entry for "Make a Good Graph" under the section for SAS Presentations at SAS Global Forum 2013, and download the examples. The code in this paper was tested using SAS® 9.3 software.

REFERENCES

Tufte, Edward. 1983. *The Visual Display of Quantitative Information*, Cheshire, CT: Graphics Press.

Cleveland, William S. 1993. *Visualizing Data*, Lafayette, IN: Hobart Press.

Robbins, Naomi. 2005*. Creating More Effective Graphs*, Hoboken, NJ: Wiley-Interscience..

RECOMMENDED READING

- *Sanjay Matange, 2012. Quick Results with SAS® ODS Graphics Designer*
- *Sanjay Matange, 2011. Tips and Tricks for Clinical Graphs using ODS Graphics*
- *Dan Heath, 2010. Creating Presentation-Quality ODS Graphics Output*

CONTACT INFORMATION

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

Name:	Sanjay Matange
Enterprise:	SAS Institute Inc.
Address:	S3014 SAS Campus Dr.
City, State ZIP:	Cary, NC 27513
Work Phone:	(919) 531- 6753
Fax:	(919) 531-4444
E-mail:	Sanjay.Matange@sas.com
Web:	http://blogs.sas.com/content/graphicallyspeaking/