

More Multivariate Data Visualization

Summary:

In our initial explorations of data visualization, we focused on a particular data set. In this reading, we consider how to write a more general solution.

Contents:

- Introduction
- Redistributing Values
- Incorporating Logarithm Calculations

Introduction

As you saw in the first laboratory on multivariate data visualization, plotting even simple data can take a series of steps as we try to figure out how to convert each data value to the range [0..300] (or whatever the width happens to be). While we did such conversion manually, it is often helpful to automate the process.

In addition to scaling values to fit on the screen, we may have to deal with distributions of data that may not scale well. We'll need to think about ways to handle such distributions.

Redistributing Values

How can we automate the process of converting a list of values to the range [0..width]? It's fairly straightforward. If we don't care about the "shifting" that we did in the lab (and some folks consider such shifting to be misleading), all we have to do is find the largest value, divide everything by that value, and then multiply by the width of the graph.

```
(define scale-values
  (lambda (values width)
    (let ((max-value (apply max values)))
      (map (lambda (value) (* width (/ value max-value))) values))))
```

If we are willing to shift the axes, we should also identify the smallest value and the difference between the smallest and largest values. We then subtract the smallest value from each value, divide by the reduced largest value, and multiply by the width of the graph. You'll have an opportunity to write such a procedure in the lab.

Incorporating Logarithm Calculations

For some distributions of data, even shifting and scaling don't seem to be enough to spread out the data. For example, consider the values (1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 5096). If we divide by the 5096, the first few values all become fairly close to 0, even when multiplied by 300.

While it may seem that such a distribution is unlikely, we do see many cases in which our values differ by many orders of magnitude. For example, the GDPs of many third-world countries are significantly smaller than those of the US. If we want to see information about both on the same graph, it is common practice to take the logarithm of the values. Such a technique results in a *log-linear* graph (if we compute logs for x values), a *log-linear* graph (if we compute logs for y values), or a *log-log* graph (if we do so for both x and y values).

When drawing such graphs, it is usually necessary to label more points on the axes to help the reader interpret the values.

Copyright © 2006 Samuel A. Rebelsky. This work is licensed under a Creative Commons Attribution-NonCommercial 2.5 License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc/2.5/> or send a letter to Creative Commons, 543 Howard Street, 5th Floor, San Francisco, California, 94105, USA.