Principles of Data Visualization

by James Bernhard

Spring 2012

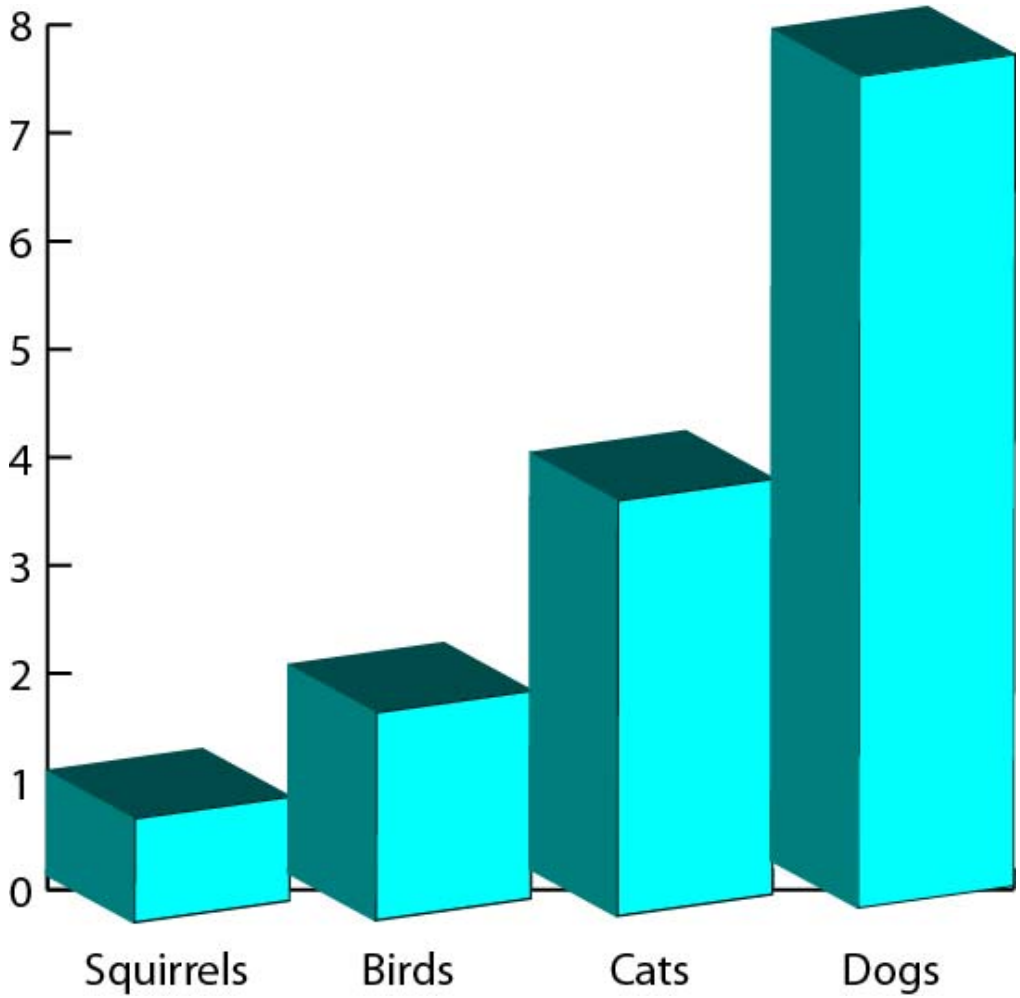
We begin with some basic ideas about data visualization from Edward Tufte (*The Visual Display of Quantitative Information (2nd ed.)*)

He gives six principles of graphical integrity

The first two are (quoted directly from page 56):

1. The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities measured.
2. Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity. Write out explanations of the data on the graphic itself. Label important events in the data.

The classic example to avoid: adding 3d where it isn't necessary



"The Lie Factor" is defined as:

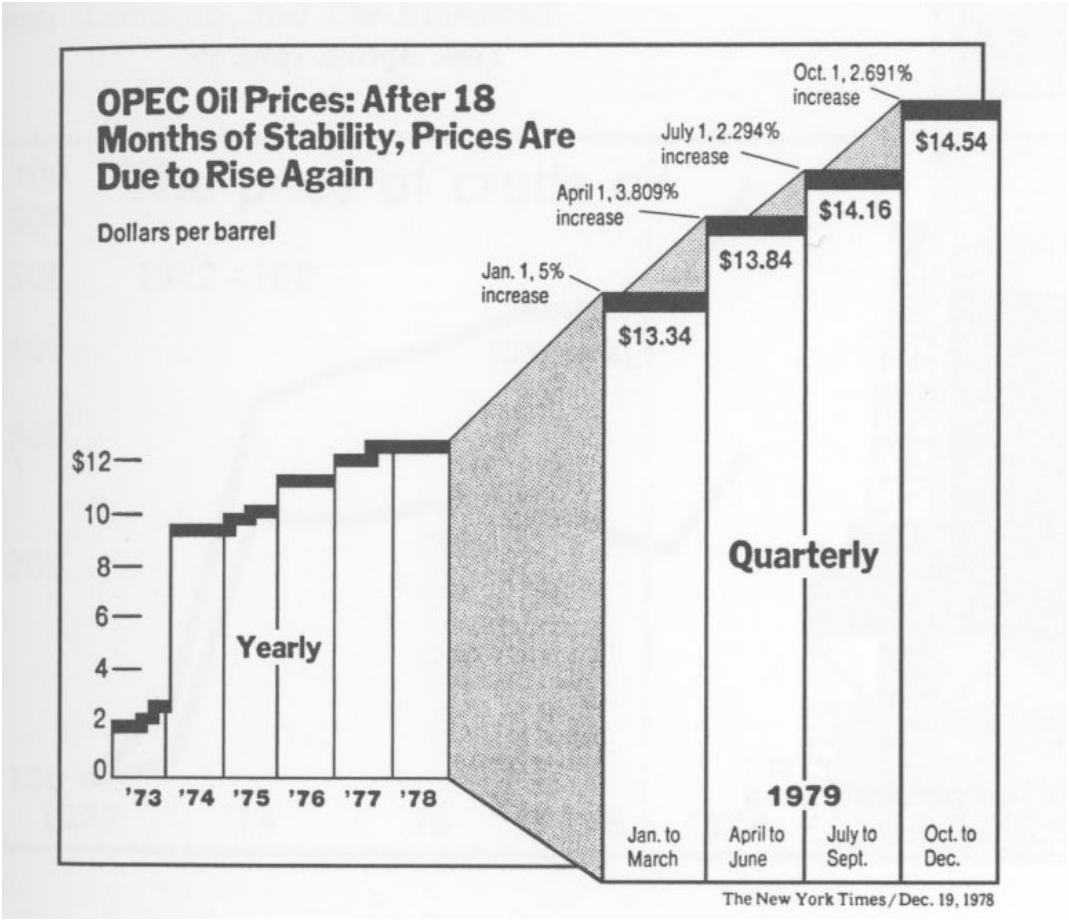
$$\text{lie factor} = \frac{\text{size of effect shown in graphic}}{\text{size of effect in data}}$$

When this isn't 1, it is not uncommon to be as large as 2 to 5

It is less usual to see a lie factor of less than 1

The third principle of graphical integrity (page 61): Show data variation, not design variation

For example, avoid things like this:



The fourth principle of graphical integrity (page 68): In time-series displays of money, deflated and standardized units of monetary measurement are nearly always better than nominal units.

The fifth (page 71): The number of information-carrying (variable) dimensions depicted should not exceed the number of dimensions in the data

I violated this principle in my gradebook visualization by using the *area* of a dot (which involves two dimenions) to indicate the weight of an assignment (which is only one variable)

However, in general it implies that you should not use 3d to represent something that only needs to be 2d, etc.

Tufte's sixth principle of graphical integrity is: Graphics must not quote data out of context (page 74).

To evaluate the information content, a useful measure (page 93) is:

$$\text{Data ink ratio} = \frac{\text{data ink}}{\text{total ink used in the graphic}}$$

This is the same as: the proportion of a graphic's ink devoted to the non-redundant display of data information

It is also the same as: 1 minus the proportion of a graphic that can be erased without the loss of data information

One of Tufte's principles for graphical excellence (page 96): Maximize the data ink ratio, within reason.

This is why Tufte frowns on heavy grid lines in the background (or even horizontal reference lines)

He extends this principle with two guidelines for how to achieve it

1. Erase non-data ink, within reason (page 96).
2. Erase redundant data ink, within reason (page 100).

To summarize Tufte's principles of how to achieve graphical excellence (page 105):

1. Above all else show the data.
2. Maximize the data-ink ratio.
3. Erase non-data ink.
4. Erase redundant data ink.
5. Revise and edit.

A useful term coined by Tufte (page 107): "*chartjunk* refers to all visual elements in charts and graphs that are not necessary to comprehend the information represented on the graph, or that distract the viewer from this information" (quoted from Wikipedia)

Common types of chartjunk:

1. Vibrating chartjunk, which is cross-hatching or other patterns that distract the mind from the information being presented
2. Grids (according to Tufte) — make them a lighter gray, not black, if you are going to use them
3. Self-promoting graphics ("The Duck"), where color schemes and patterns are introduced for artistic appeal rather than information content.

Tufte gives one other useful pair of principles on the aspect ratio of a display:

- If the nature of the data suggests the shape of the graphic, follow that suggestion.
- Otherwise, move toward horizontal graphics about 50 percent wider than tall (approximately a 3:2 aspect ratio).

We now describe some principles given by William S. Cleveland in his book *The Elements of Graphing Data* (1985)

Many (but not all) of these are quite similar to those of Tufte, although his work in this area preceded Tufte's work

His four major categories in the principles of graph construction:

- Clear vision
- Clear understanding
- Scales
- General strategy

Clear vision

- Make the data stand out. Avoid superfluidity.
- Use visually prominent graphical elements to show the data.
- Do not clutter the data region.
- Use a reference line when there is an important value that must be seen across the entire graph, but do not let the line interfere with the data.
- Do not allow data labels in the data region to interfere with the quantitative data or to clutter the graph.

- Avoid putting notes, keys, and markers in the data region. Put keys and markers just outside the data region and put notes in the legend or in the text.
- Overlapping plotting symbols must be visually distinguishable.
- Superposed data sets must be readily visually discriminated.
- Visual clarity must be preserved under reduction and reproduction.

Clear understanding

- Put major conclusions into graphical form. Make legends comprehensive and informative.
- Error bars should be clearly explained.
- Proofread graphs.
- Strive for clarity.

Scales

- Choose the range of the tick marks to include or nearly include the range of data.
- Subject to the constraints that scales have, choose the scales so that the data fill up as much of the region as possible.
- It is sometimes helpful to use the pair of scale lines for a variable to show two different scales.
- Choose appropriate scales when graphs are compared.
- Do not insist that zero always be included on a scale showing magnitude.
- Use a logarithmic scale when it is important to understand percent change or multiplicative factors.
- Showing data on a logarithmic scale can improve resolution.

General strategy

- A large amount of quantitative information can be packed into a small region.
- Graphing data should be an interative, experimental process.
- Graph data two or more times when it is needed.
- Many useful graphs require careful, detailed study.