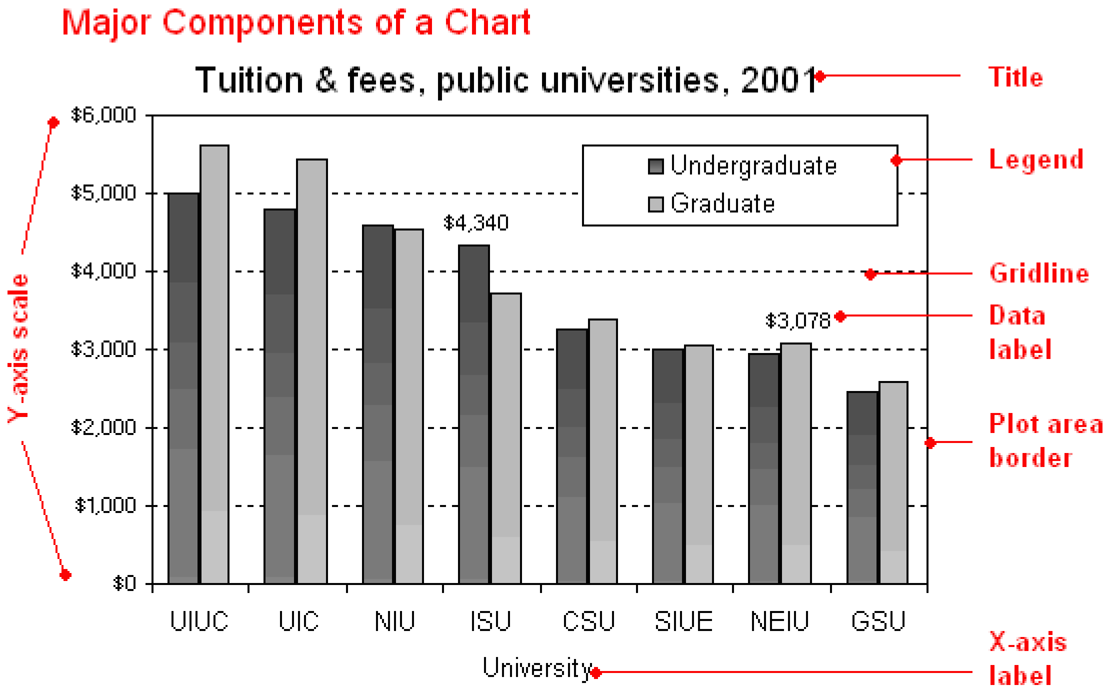
Presenting Data: Tabular and graphic display of social indicators

(under construction)

Gary Klass
Illinois State University

Home | Numbers? | Good Tables | ■ Good Charts
References | Course Page | Chart of the week

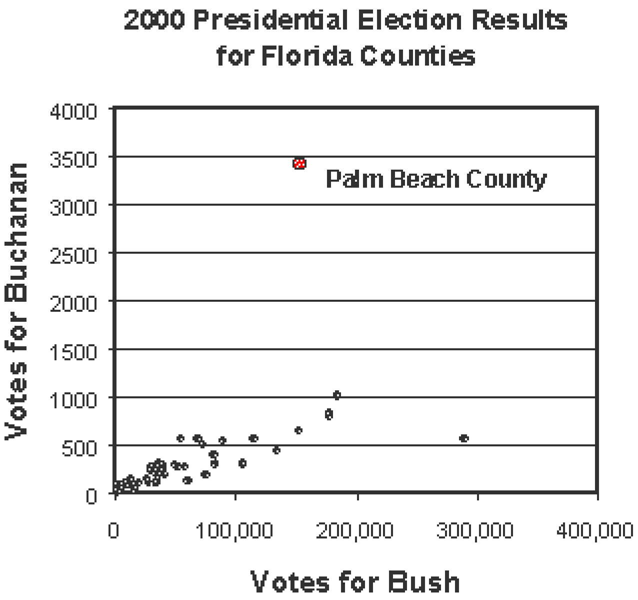
Constructing Good Charts and Graphs



General Principles of Graphic Display

A graphical chart provides a visual display of data that otherwise would be presented in a table; a table, one that would otherwise be presented in text. Ideally, a chart should be able to convey to the reader ideas about the data that would not be as readily apparent if they were displayed in a table or as text.

Figure 1.



source: Prof. Greg D. Adams, Carnegie Mellon University <gadams@andrew.cmu.edu> and Prof. Chris Fastnow, Chatham College <cfastnow@chatham.edu>

As an example, consider one of the many disputes of the 2000 Presidential election concerned the "butterfly" ballot used in Florida's Palm Beach County. Many Democrats alleged that voters who had intended to vote for Gore were tricked by a poorly designed ballot into voting for Patrick Buchanan instead. To see what the dispute was about, consider this -- exaggerated -- interpretation of how the ballot looked to voters.

One way to estimate how many votes Gore lost due to the ballot design is to try to determine how many extra votes Buchanan received. Figure 1 illustrates one of many approaches to this; each data point on the scatterplot represents the number of votes for Bush and Buchanan for each of Florida's 67 counties.

Although there are problems with these data (using the percentage vote for the candidates, Buchanan's vote does not appear as exceptional), the figure makes a point that would not be readily apparent from even an extended viewing of the raw vote totals,

The three standards for tabular display of data -- the **efficient** display of **meaningful** and **unambiguous** data -- apply to charts and graphs as well. As with tables, it is crucial to good charting to choose meaningful data, to clearly define what the numbers represent, and to present the data in a manner that allows the reader to quickly grasp what the data mean.

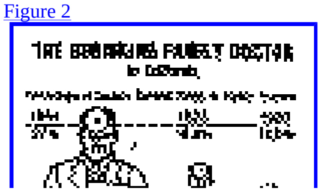
As with tabular display, data ambiguity in charts arises from the failure to precisely define just what the data represent. Every dot on a scatterplot, every point on a time series line, every bar on a bar chart represents a number (actually, in the case of a scatterplot, two numbers). It is the job of the text on the chart to tell us just what each of those numbers represents. If a number represented in a chart is, say, 33½, the text in the graph -- in the title, the axis labels, the data labels, the legend, and sometimes the footnote -- must answer question: "Thirty-three and a half what?"

Good graphical display, however, requires more than good tabular display as it draws on the talents of both the scientist and the artist. You have to know and understand your data; but you also need a good sense of how the reader will visualize the graphical elements of the chart.

Good graphical display will tell the reader things about the data that are no apparent when the data are displayed in tabular form. Bad graphical display, on the other hand, either distorts or hides what the data would tell the reader. Data distortion occurs when the graphical elements of a chart do not give a true picture of what is going on with the data. Data distraction occurs when the graphical elements of the chart distract the readers from seeing what the data might tell them.

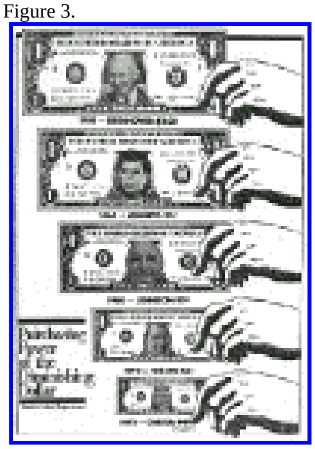
Data Distortion.

Figure 2



Before the development of spreadsheet graphing, the most common graphical mistake was the use of artist-drawn 3-D images with the height of 3-D objects representing the magnitude of the data points. In these charts, both the height and the width of the drawn object increase proportionate to the magnitude of the data points. The effect is to exaggerate the differences in magnitude as the viewer tends to perceive the area of the figures rather than just the height as representing the magnitude. The incredible shrinking family doctor (shown in Tufte, p. 69) is a classic example. In this chart the 1990 doctor is a bit less than half the height of the 1964 doctor. Each doctor has the same relative shape. Image two doctors with the same average physical shape, one less than 4 feet tall, the other 8 feet tall. If the 4 ft. doctor weighed 100lbs., how much would the 8 ft. doctor weigh? Certainly much more than 200lbs.

Figure 3.



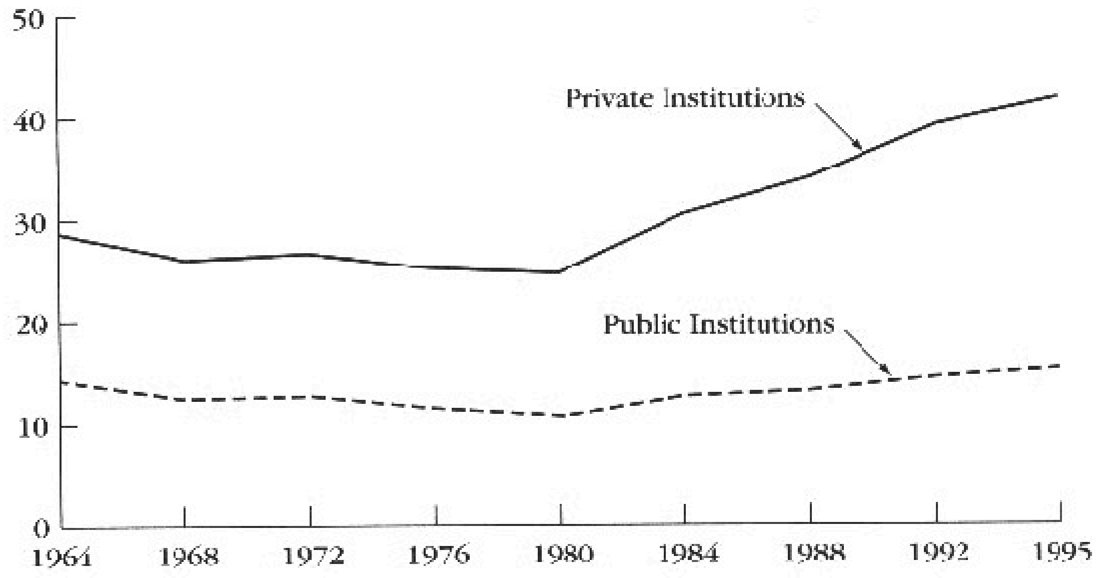
In the figure on the left, the width of the dollar bills is used to represent the declining values of the US dollar from $1.00 in 1958 to 46 cents in 1973. Essentially, in 1973 it took a bit more than 2 dollars to buy what one dollar bought in 1958. But the graphic designer here has reduced both the width and the height of the dollar bill. In terms of area, the 1958 dollar is five times larger than the 1973 dollar. The distortion, measured by what Edward Tufte calls the lie factor (the ratio of the size of the effect shown in the graph to the size of the effect in the real data) is 5 to 2.

With the development of spreadsheet graphics, such visual distortions are no longer common, and the Art of Lying with graphics has become a technology rather than an art. Today, altogether new forms of bad graphical design predominate.

We will consider now some of the more complicated ways of using graphical display to mislead. Figure 4 is a time series chart originally printed in a public policy textbook authored by four professors of political science employed by three public universities.

Figure 4.

FIGURE 9-12 ▪ Average Tuition, Room, and Board as a Percentage of Median Family Income, 1964–1995



SOURCE: U.S. Dept. of Education.

source: Cochran, 347

Their interpretation of the data is as follows:

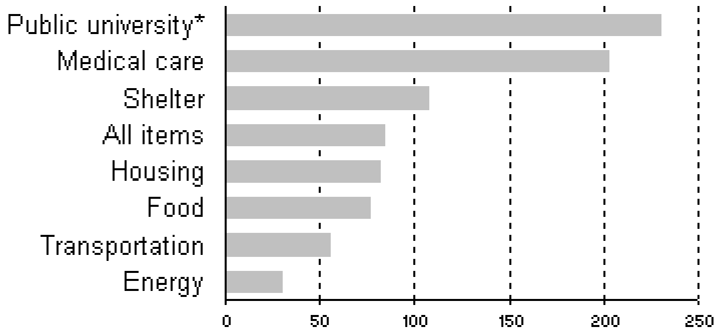
> There is some evidence that the cost of higher education may not have escalated so much... Figure 9-12 *[i.e. figure 4, above]* reflect the average cost for tuition, room, and board as a percentage of median family income from 1964 to 1995. While private institutions have increased costs substantially, public university costs have remained constant. This indicates that the increased costs associated with higher education may be quite reasonable

> when compared to family income levels. (Cochran 346-7)

Note the ways in which the authors have understated the rising costs of public university education. First, the costs are "deflated" using median family income rather than the more conventional consumer price index. Especially for the years after 1982, median family income rose much faster than the consumer price index. Second, graphing both the private and public data on the same graph enlarges the scale on which the public data is displayed. It's hard to tell from the graph, but between 1980 and 1995 it appears that public university costs increased from around 11% of family income to near 15% -- in effect the share of family income going to public university costs has increased by a third. The third way of minimizing the cost increases that have occurred since 1980 is to extend the time series back to 1965.

Figure 5.

% Increase in Prices, by Economic Sector, 1981-1999



*Public 4-year universities, tuition, room and board
source:US Statistical Abstract 2000 CD-Rom, table 311

A completely different picture emerges if one were to compare the rate of increase in public university costs to the rate of increases in other sectors of the economy. On the left, we see that from 1981 to 1999 -- over the lifetime of today's college student -- public university costs have risen faster than any other sector of the economy. Faster even than rising medical care costs. In addressing the topic of health care inflation, the same authors note that: "Cost escalation in the medical field has been constant," and spend four pages of text addressing the reasons for the increases. (pp. 268-72).

Examine this chart from the UNICEF that purports to demonstrate that the gap between rich and poor countries is increasing. We can see that the per capita GNP of the wealthiest countries has slightly almost doubled (from about $12,000 to about $26,000), but it is not clear that the GNP hasn't doubled or tripled among either the middle or low-income countries.

Here's an example from the same source that seems to distort the data. Note the size of the two arrows, but look carefully at the first arrow -- the negative $18 change is represented not by the arrow, but by the little line below it.

Data Distractions:

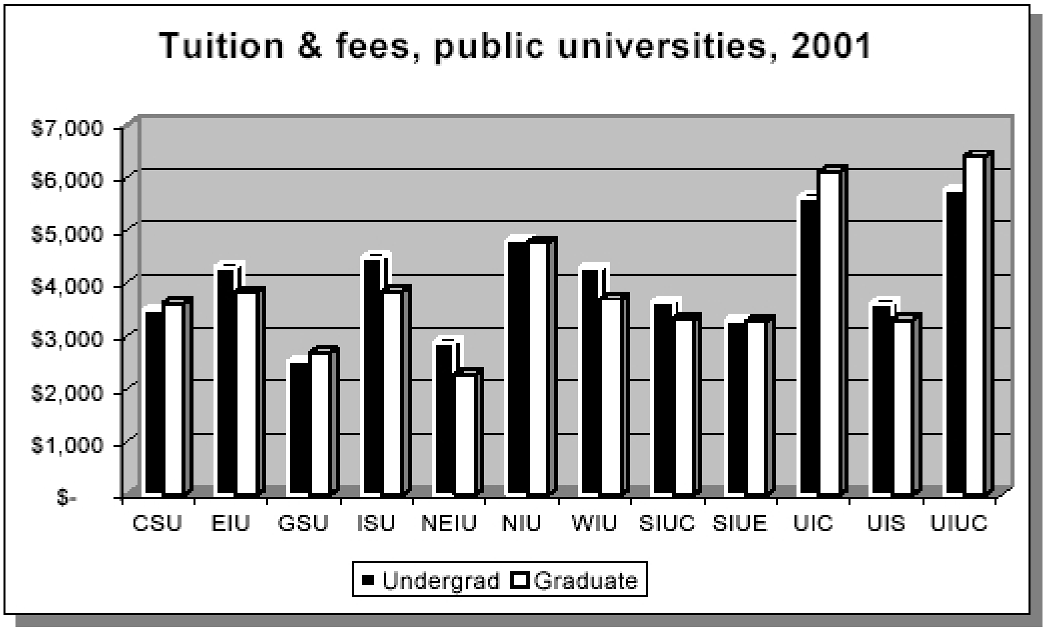
Edward Tufte's fundamental rule of efficient graphical design is to **minimize the ratio of ink-to-data**. This is essentially the same advice offered by Strunk and White to would-be writers:

> "A sentence should contain no unnecessary words, a paragraph no unnecessary sentences for the same reason that a drawing should contain no unnecessary lines and a machine no unnecessary parts." (23)

The primary causes of extraneous lines in charting graphics today are the 3-D options offered by conventional spreadsheet charting software. These 3-D options serve no useful purpose; they add only ink to the chart, and more often than not make it more difficult to estimate the values represented. Even worse are the spreadsheet options that allow one to rotate the perspective. For those who would take bad graphical display to even higher levels, the Excel spreadsheet program offers the option of doughnut, radar, cylinder, cone, bubble charts. All these are de facto violations of the Tufte's fundamental rule.

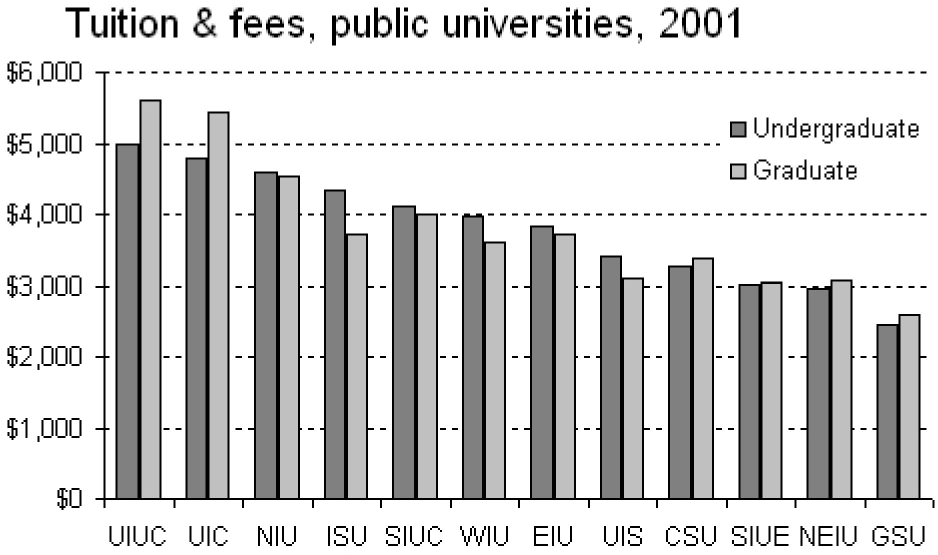
As was the case with tabular display, sorting data by the most meaningful variable greatly aids in the interpretation of data. A common charting mistake is to sort the data alphabetically.

Figure 6a



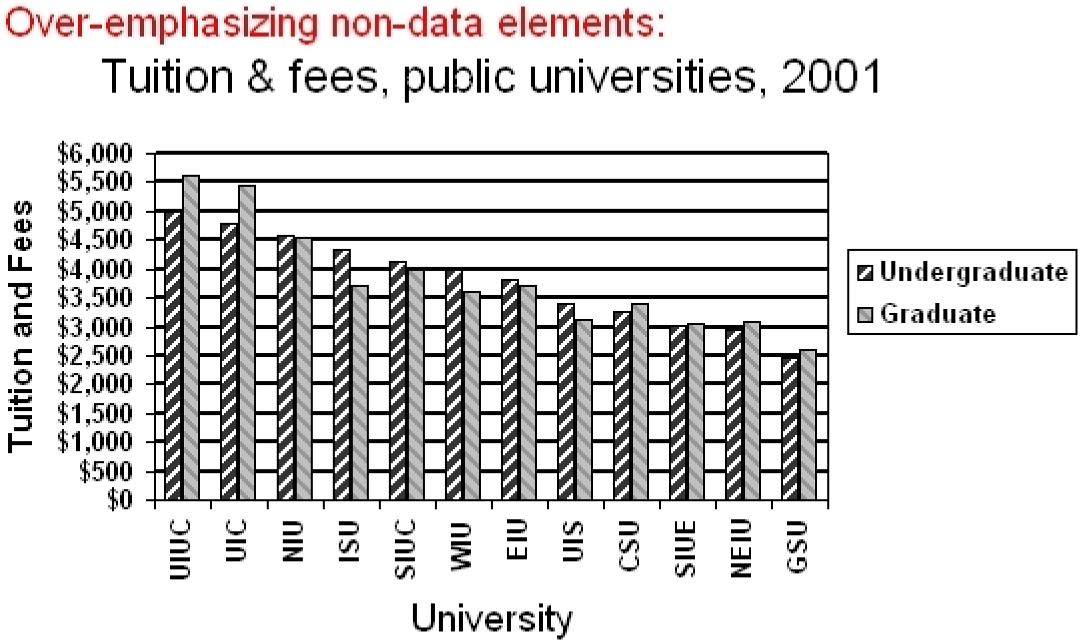
source: IBHE Annual Report

Figure 6b: revised chart.



The corollary of Tufte's fundamental rule is "Show the Data". To often, unimportant, unnecessary, or irrelevant graphical and textual elements dominant charts, distracting from the chart's fundamental purpose of displaying data. A good chart maximizes the presentation of data and minimizes the display of non-data elements. In the preceding chart, removing the 3-D effect, placing the legend inside the plot area permits a longer Y-axis, de-emphasizing the gridlines all serve to enhance the display of the differences in the relative length of the bars. In the following chart, many of these graphical elements are over-emphasized:

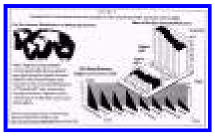
Figure 6c:


Over-emphasizing non-data elements:

Tuition & fees, public universities, 2001

One way to minimize the ink-to-data ratio is to show more data. Too much data on a chart -- too many trendlines in a times series chart, for example -- can overwhelm the data presentation, but the more common mistake is to use several charts to display data that could just as easily be displayed in one. The tuition and fee charts above could have been separated into graduate and undergraduate charts or, worse, as 12 separate charts for each university. Avoid, if at all possible, forcing the reader to make comparisons between to charts.

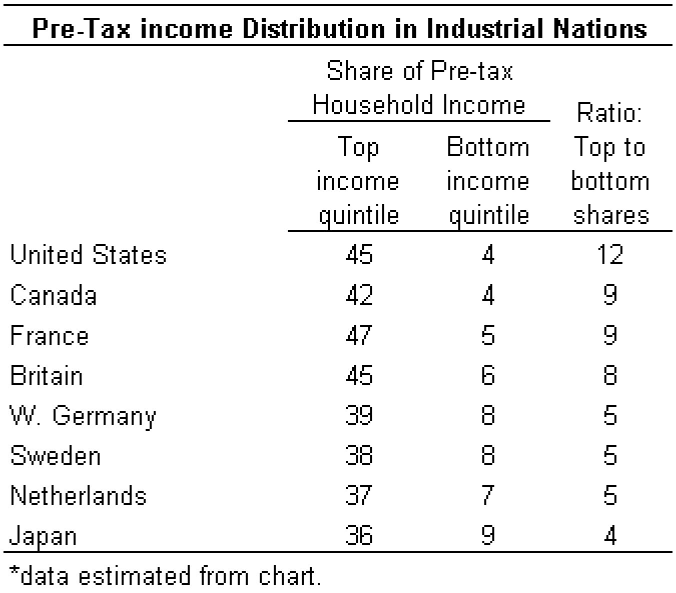
ChartJunk. Not content with the distractions and distortions made possible by the use of 3-D effects, charters sometimes feel the need to add all sorts of other "Chartjunk" to a graph. In the graphics on the left, Kevin Phillips (1991, 9) is trying to make the point that income is more inequitably distributed in the United States than in other countries (click on the image for a full display).



Note the extraneous features of this in this graphic.

- A completely irrelevant map of the world.
- Two entirely different kinds of 3-D charts displayed at two different perspectives.
- Country names are repeated three times.
- To display 24 numeric data points, 28 numbers are used to define the scales.
- The countries are sorted in no apparent order (not even alphabetically).
- Note the use of the letter " I " to separate the countries on the bottom chart.

While it might be possible to display these data better graphically, a table does the job quite nicely:



Pre-Tax income Distribution in Industrial Nations

	Share of Pre-tax Household Income		Ratio:
	Top income quintile	Bottom income quintile	Top to bottom shares
United States	45	4	12
Canada	42	4	9
France	47	5	9
Britain	45	6	8
W. Germany	39	8	5
Sweden	38	8	5
Netherlands	37	7	5
Japan	36	9	4

*data estimated from chart.

Other Chartjunk examples

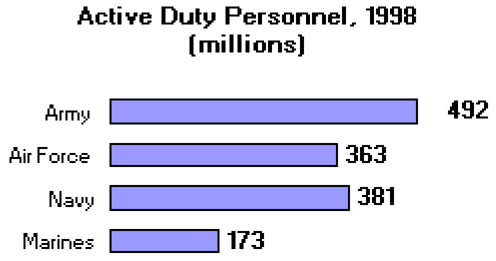
Other Bad Charting - Examples

Types of Charts

Most charts are a variation on one of four basic types: Scatterplot, time series chart, bar chart and pie chart; most data display charts are one of many variations or combinations of these types. Choosing the right type of chart depends on the characteristics of the data and the relationships you want displayed.

With more than a few data points, Scatterplots and time series charts, can display patterns of relationship among two or more (with time series charts) variables that would not be obvious in a tabular displays. Both scatterplots and time series charts require at least two interval-level variables.

Bar charts and pie charts are used to display data for categorical variables (although a Bar chart is sometimes also used in a time series chart).



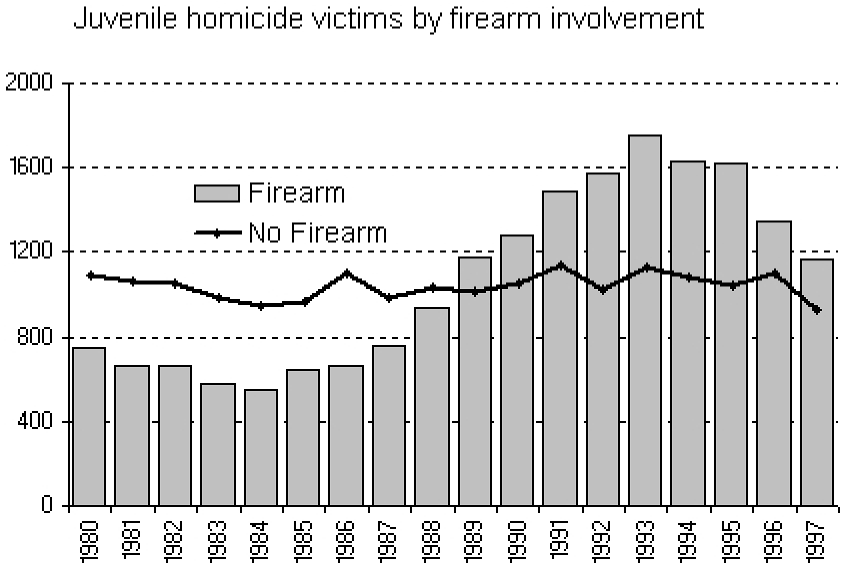
Bar charts, especially one-variable bar charts, are generally not an efficient means of displaying data. The bar chart on the left, for example, contains an absolute minimum amount of ink; the gridlines, plot area frame, and axes have all be removed. Still, the bars themselves are unnecessary -- serving only to provide a graphical representation of the size of the four numbers. Because the size of each bar in a bar chart represents a number that would otherwise be displayed in a table, one could justify the use of bar charts only if one assumes that the reader would more easily discern the relative size of two or more number if they are represented graphically.

Pie charts are even worse in this respect than bar charts. Any data that can be displayed in a pie chart can also be displayed in a bar chart.

Times Series Charts:

The annual time series chart is one of the most efficient means of displaying large amounts of data in ways that provide for meaningful analysis. .

Figure 2



source: Office of Juvenile Justice Delinquency and Prevention, *Statistical Briefing Book*.
http://ojjdp.ncjrs.org/ojstatbb/html/qa128.html
(viewed on 7/21/01).

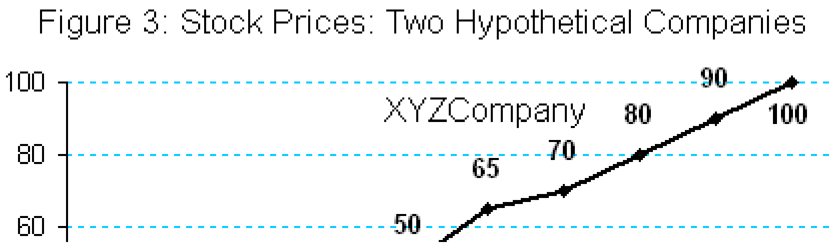
Standards for Times Series Charts:

- Time is almost always displayed on the X-axis from left to right.
- Display as much data with as little ink as possible.
- Avoid clutter
- In bar charts using time as a category variable on the X-axis, always display the time points on adjacent bars.
- Make sure the reader can clearly distinguish the lines for separate variables.
- When displaying fiscal or monetary data over-time, it is often best to use deflated (i.e. adjusted for inflation, per capita or "as a percent of GNP") measurements.

Problems with Time Series charts:

Scaling: When two variables with numbers of different magnitudes are graphed on the same chart, the variable with the large scale will generally appear to have a greater degree of variation; the smaller-scale variable will appear relatively "flat" even though the percentage change is the same.

In figure 3, XYZCompany's stock seems to be growing much faster than ABCorp's, yet the rate of increase is identical.



When the differences
in scale are so great as to eliminate most of the perceived variation in the smaller-scale variable, using a second scale (displayed on the right-hand side) is sometimes preferable, although this makes the interpretation of the graph more complicated. In figure 4a and 4b, the difference in scales for the poverty and unemployment are almost, or perhaps no quite, great enough to justify a second axis scale.

Figure 4a: Poverty and Unemployment 1959-00

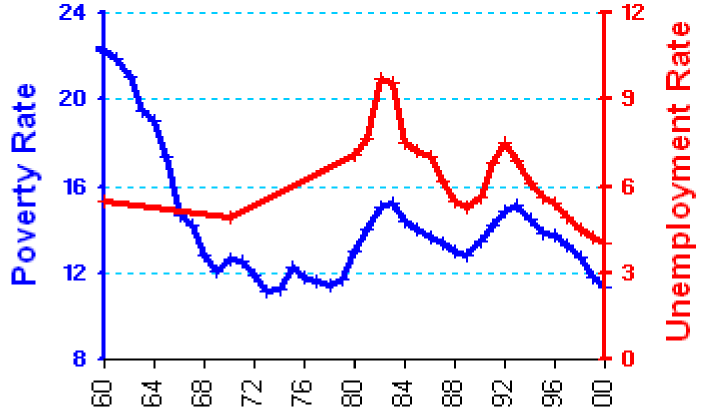
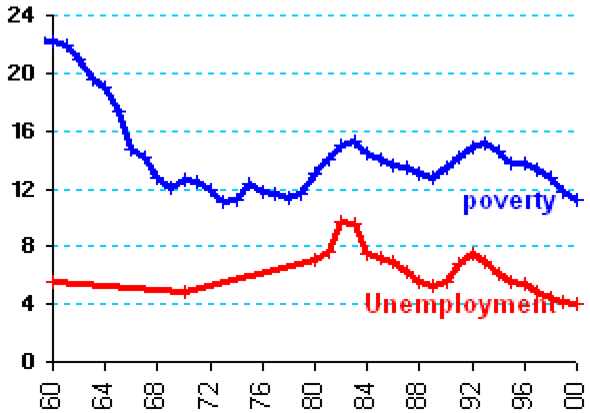
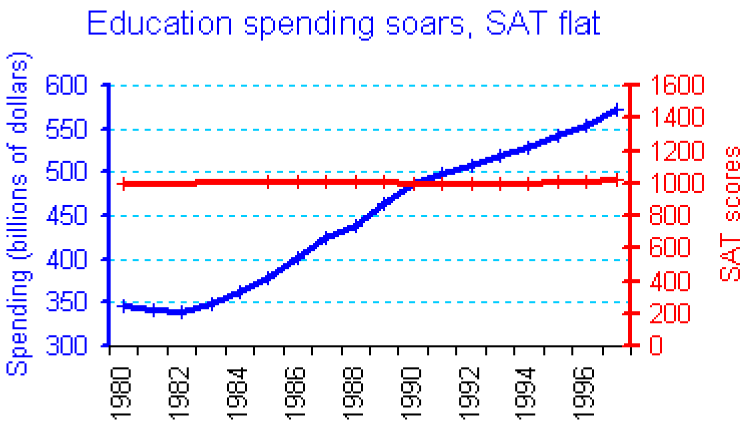
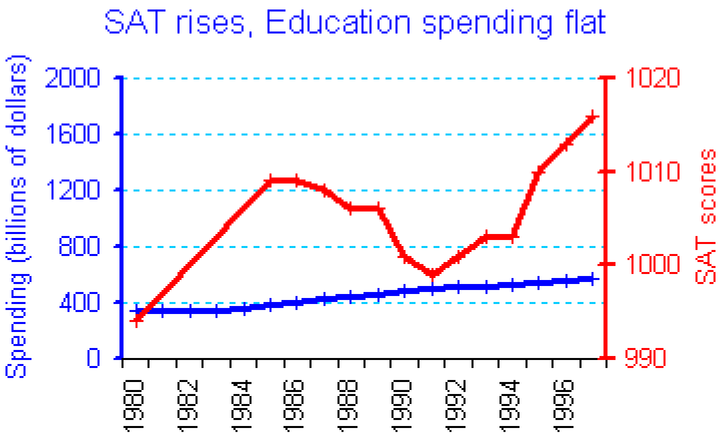


Figure 4b: Poverty and Unemployment 1959-00

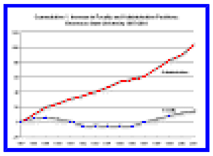


Many who have written about graphical distortion condemn the use of two-scale charts because the relative size of the two scales is completely arbitrary. This is true. Increasing the size of either scale in figure 4 will make the corresponding line appear flatter. One the other hand, using the same scale to measure variables of different magnitudes, as we saw in figure 3, has problems too.



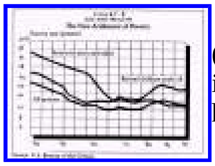


Another solution is to rescale the variables. This is often done with price or other monetary data by setting the initial year values of both variables equal to 100. Multiply each variable by the 100 times the inverse of the value for the base year.



Note that a similar problem may also occur with bar charts.

Crossing lines. When several times series lines are printed in black and white, it is sometimes difficult to separate out the different tend lines. Mixing solid, dotted, and dashed lines for each variable may solve this problem, although it is sometimes difficult to distinguish between dotted and dashed lines. Here's an example of a published chart with ambiguous lines.

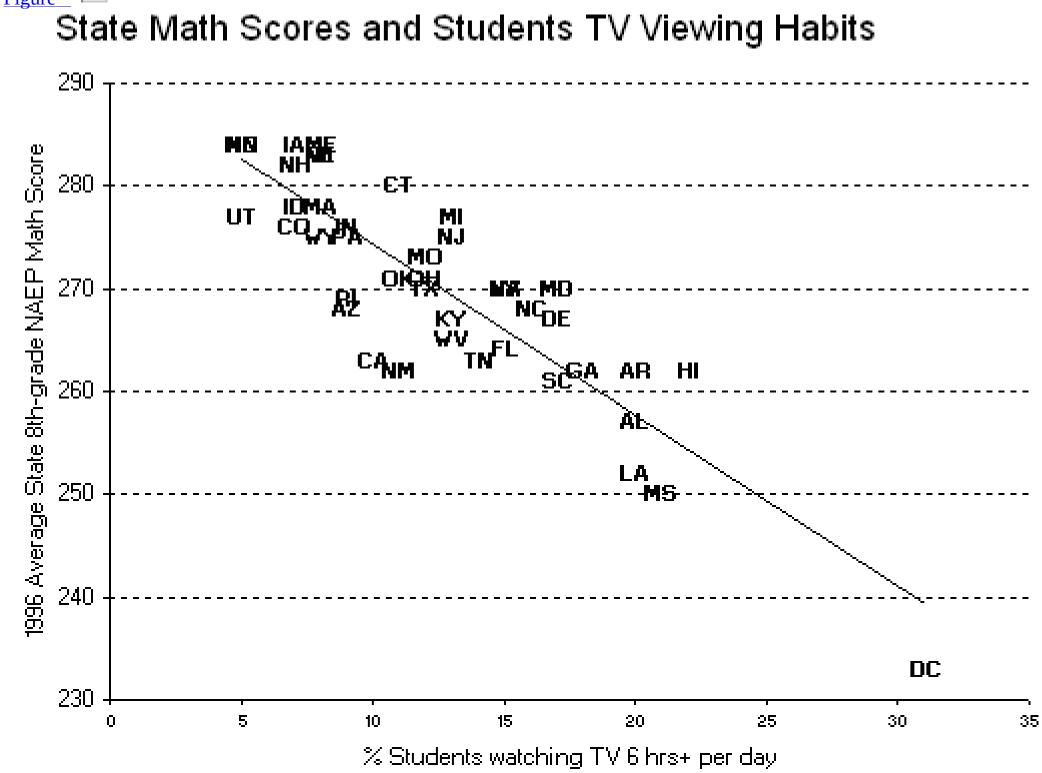


(click on thumbnail for larger image, close new window to return here)

Scatterplots.

The two-dimensional scatterplot is the most efficient medium for the graphical display of data. A simple scatterplot will tell you more about the relationship between two interval-level variables than any other method of presenting or summarizing such data.

Figure



With good labeling of the variables and cases and common sense scaling of the X and Y-axes, there's not a lot that can go wrong with a scatterplot.

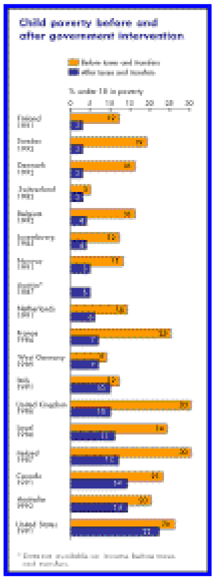
Basic rules of Scatterplotting:

- Have two interval-level variables.
- Use the titles and axis labels to clearly define the variables.
- Be sure to define the cases (e.g., indicate whether the data points represent persons, cities, or countries).
- Scale each axis to each variable's full data range.
- If there is an implied causal relationship between the variables, the independent variable (the one that causes the other) should be on the X-axis and the dependent variable (the one that may be caused by the other) should be on the Y-axis.

In the "Math and TV viewing" chart, above, TV viewing is the independent variable. (If you were trying to predict which types of students watch the most TV, the axes would be reversed.) The scatterplot contains two optional plotting features: a regression trendline denoting the linear relationship between the two variables and the use of State postal ID data labels to indicate each state's position on the chart (these labels require a special add-in to the Excel program). Although the chart suffers from overlapping data labels, the interpretation is straightforward; the higher the percentage of students in a state watching more than 6 hours of TV each day, the lower the state's math scores.

Bar Charts:

Time series (and scatterplot) graphs are often an efficient means of displaying lots of data in little space and depicting relationships that are not easily discerned by the viewer were the data presented in tabular format. Bar charts, however, often contain little data, a lot of ink, and rarely reveal ideas that cannot be presented much more simply in a table. (Note: in this discussion both vertical and horizontal bar charts are called bar charts; MS Excel refers to horizontal bar charts as column charts.)

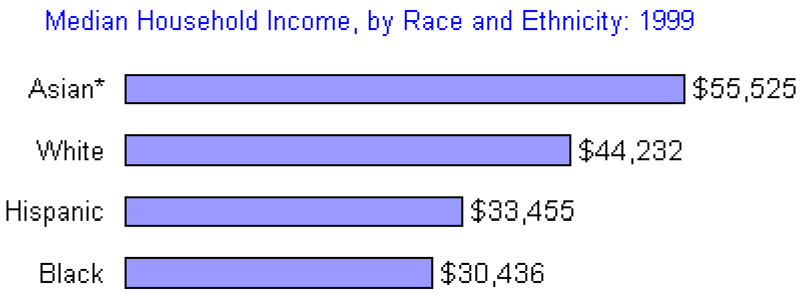


The child-poverty chart on the left (click on it for a full view) is an example of bar-charting at its best. For a bar chart, it contains a lot of data, depicting 35 data points for the two main variables and 18 cases. The cases are sorted on the most meaningful variable. Gridlines are made unnecessary by displaying the actual data points (although the 0 to 30 scale at the top is also unnecessary).

Most importantly, the data provide for some really interesting comparisons. Comparing the blue bars tells us that the United States has the highest child poverty rate among developed nations; comparing the blue with the orange we see that United States government policies do the little to address the problem of child poverty. One does need to know that "transfers" consist mostly of government social welfare payments; the orange bar represents the percentage of children that would have been in poverty without such income supports.

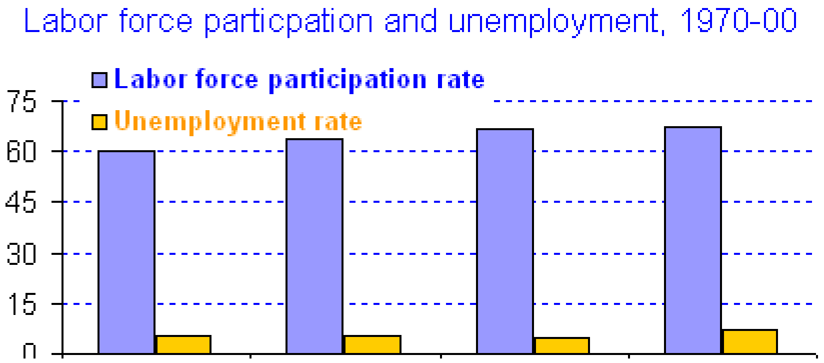
Look at this chart and you can quickly grasp the main points, but then spend some time with it and you'll discover other interesting things. Note, for example, that the top four countries are located in Northern Europe; the six countries at the bottom are either European-island or non-European nations.

Nevertheless, this bar chart does little more than a table would: indeed, it presents all the numbers and text that would be included in a table, adding only the blue and orange bars. In effect, the only function of the bar chart's bars is to provide the reader with a visual idea of how big the numbers are.



*Asan or Pacific Islander
Bureau of Labor Statistics:
http://www.bls.census.gov/cps/ads/2001/susenote.htm

Minimizing the ink-to-data ratio is especially important in the case of bar charts. The chart on the left features the absolute minimal amount of ink while still retaining both the bar chart format and a good definition of the data. Never use a 3-D bar chart. Keep the gridlines faint. Display no less than four or more than six numbers on the y-axis scale. If there are fewer than five bars, consider using data labels rather than a y-axis scale; it doesn't make sense to use a five-numbered scale when the exact values can be shown with four numbers. Microsoft Excel does allow stacked and stretched images on bars, this too falls into the category of distraction gimmicks.

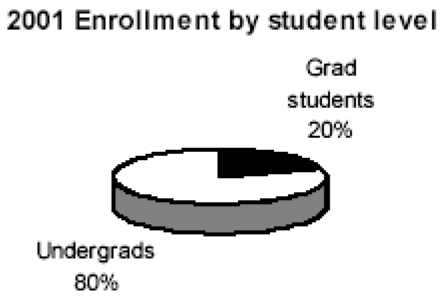


As we see in the chart on the right, when more than one variable is bar-charted, scaling is as big a problem as it is with times series charts.

Color is of greater

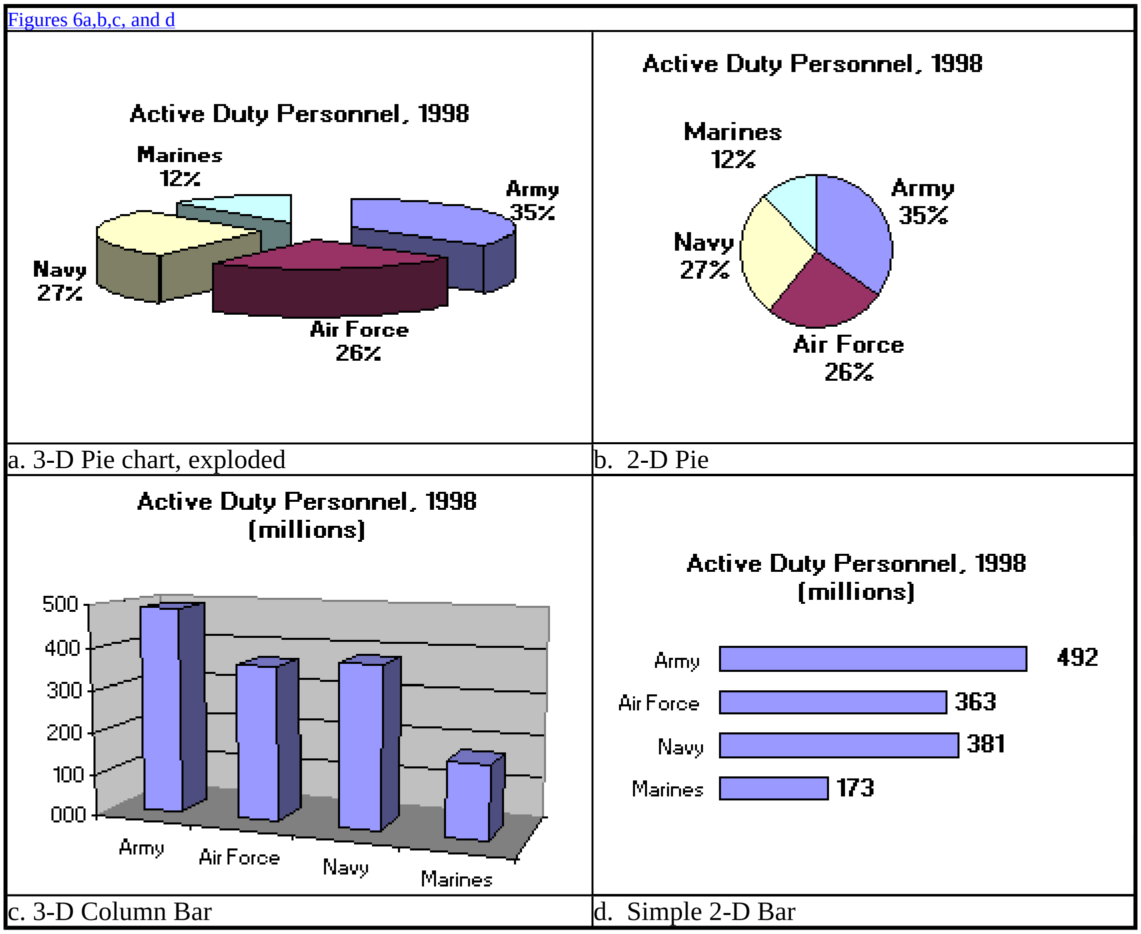
consideration in
bar charts than in line or scatterplots. Softer colors are better than primary colors (which I tend to over use on line charts).

Pie Charts



Pie charts should rarely be used. Pie charts usually contain more ink than data and provide for a poor representation of the magnitude of the data points. It is more difficult for the eye to discern the relative size of pie slices than it is to assess relative bar length. With pies, it is difficult to figure out whether the Navy or Air Force is larger without looking at the numbers; from the bar charts it is obvious. 3-D pie charts are even worse, as they also add a visual distortion (in this case, the thick 3-D band exaggerates the size of the undergraduate slice).

Figures 6a,b,c, and d



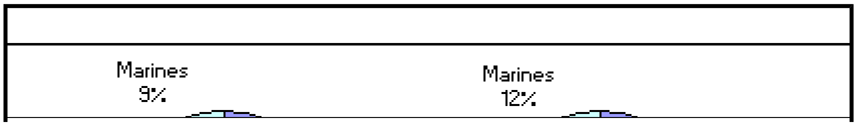
a. 3-D Pie chart, exploded

b. 2-D Pie

c. 3-D Column Bar

d. Simple 2-D Bar

Note how much less ink the 2-D bar charts uses compared to the 3-D bar. Using data labels rather than a y-axis scale in this case reduces the number of numbers displayed from 6 to 4, and adds precision as well. Normally, I would have sorted the data here, so that the Navy would be between the Army and Air Force, but since the Marines are a part of the Navy (and the Air Force, originally, a part of the Army), this order made more sense. A strict application of the ink-to-data in this case, however, would eliminate the bars altogether and simply present the data as a table.



Two Pies are worse than one: Pies are even less effective when an additional variable is added and comparisons between pies are required (sometimes by adjusting the relative size of the pies).

Tips on Using MS Excel to Prepare Charts and Graphs.

References:

editing notes:

In addition, graphs should display as much data as can by quickly interpreted with as little ink as possible.

- *Minimize the ink-to-data ratio*. Avoid chart junk, unnecessary 3-D effects, excessive gridlines. Avoid redundant labeling.
- Use only the data necessary to display the relationships.
- Display Multivariate rather than univariate relationships.
- Highlight the most interesting relationships.
- Avoid distorting the data.
- Use consistent formatting and style across a set of graphs.
- Clearly define the data with titles, labels, legends and notes.

$88,000,000
July 18
$54,712,749
Redefining normal
The Campaign for Illinois State University