

**STAT 6560**  
**Graphical Methods**  
**Spring Semester 2011**

**Dr. Jürgen Symanzik**

Utah State University

Department of Mathematics and Statistics

3900 Old Main Hill

Logan, UT 84322-3900

Tel.: (435) 797-0696

FAX: (435) 797-1822

e-mail: [symanzik@math.usu.edu](mailto:symanzik@math.usu.edu)

Web: <http://www.math.usu.edu/~symanzik/>



# Contents

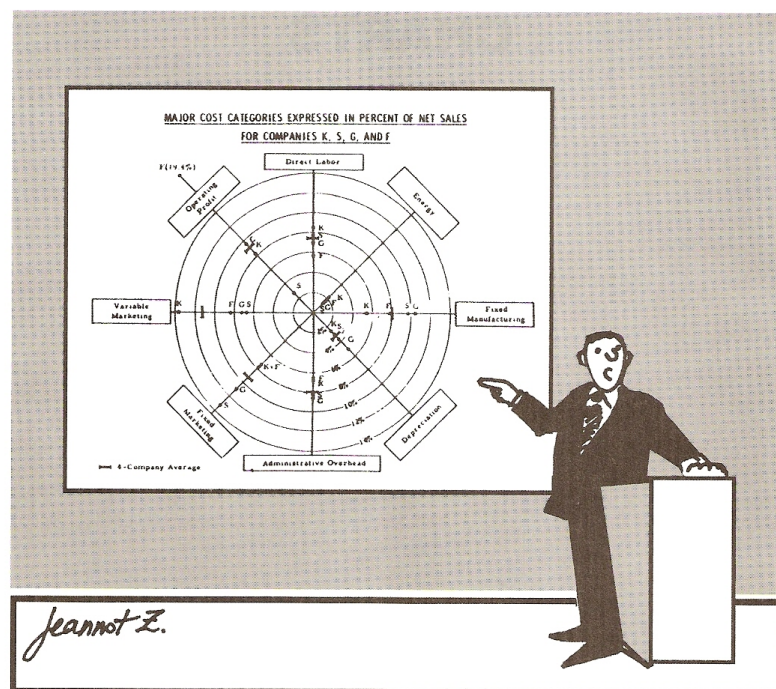
<b>Acknowledgements</b>	<b>1</b>
<b>1 Introduction: A Couple of Good and Bad Examples</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.2 Why Graphics !? . . . . .	2
1.3 How to Display Data Badly . . . . .	3
1.3.1 Don't show much data . . . . .	4
1.3.2 Show the data inaccurately . . . . .	9
1.3.3 Obfuscate the data . . . . .	16
1.4 Bad Graphics are Everywhere — In Space and in Time . . . . .	32
1.5 Rules for Good Data Displays . . . . .	43
1.6 Further Reading . . . . .	47
<b>2 History of Statistical Graphics: Plots, People, and Events</b>	<b>49</b>
2.1 General History . . . . .	49
2.1.1 Milestones in the History of Data Visualization (According to Friendly) . . . . .	50
2.2 Selected People . . . . .	51
2.3 Statistical Graphics and Events in History . . . . .	64
2.3.1 John Snow and the Cholera Epidemic in London, 1854 . . . . .	64
2.3.2 The Challenger Disaster, 1986 . . . . .	68
2.4 Further Reading . . . . .	72
<b>Appendix</b>	<b>73</b>
<b>Homework Assignments</b>	<b>74</b>
<b>Homework Assignment 1</b>	<b>1</b>
<b>Homework Assignment 2</b>	<b>1</b>
<b>References</b>	<b>4</b>

# Acknowledgements

This course uses some of the course materials provided by Dr. Mike Minnotte (formerly USU, now with the University of North Dakota) as held in the Fall 2006 semester. Additional material have been taken from other Statistical Graphics courses, such as the ones offered by Dr. Di Cook (Iowa State University: <http://www.public.iastate.edu/~dicook/>) and Dr. Dan Carr (George Mason University: <http://mason.gmu.edu/~dcarr/>). We are likely to include parts from additional Web sources that will be specified later on.

Thanks are also due to the seven students who took Stat 6560 with me in the Spring 2009 semester for their valuable comments that helped to improve and correct these lecture notes.

Jürgen Symanzik, January 11, 2011.



"What do you mean, what does it mean?"

Figure 1: Zelazny (2001), p. x, Cartoon.



# 1 Introduction: A Couple of Good and Bad Examples

(Based on Wainer (1997), Chapter 1 & Tufte (1983), Chapter 2)

## 1.1 Motivation

Statistical graphics and data visualization are critical elements of modern data analysis and presentation. From initial exploration of a data set to the final presentation of results to the end user, statistical graphics play a vital role in shaping our understanding of our data. Through proper use of graphics, we can make critical discoveries, and communicate them clearly. Conversely, poor use or misuse of graphics can seriously mislead (by accident or design).

In this course, we will start with presentation graphics, including discussion of both tools and principles which lead to clear communication and those which serve only to confuse or mislead. We will spend most of the semester in exploratory graphics and data analysis, including data mining. This will be broken down largely by the dimension of the applicable data. One- and two-dimensional datasets require and allow far different methods than those of more than three dimensions. Categorical and regression data call for their own specialized methods.

Even more than most aspects of statistics, graphics and visualization involve art as well as science. In most cases, there are many reasonable approaches. Only an understanding of the options available and the underlying principles will lead to a successful analysis and presentation.

Via graphics, otherwise boring statistical information can become really exciting and entertaining. Hans Rosling ([http://en.wikipedia.org/wiki/Hans\\_Rosling](http://en.wikipedia.org/wiki/Hans_Rosling)), a Swedish medical doctor and statistician, “sells” statistics extremely well. Enjoy his talk *Debunking Third-World Myths with the Best Stats You’ve Ever Seen*, accessible at [http://www.ted.com/index.php/talks/hans\\_rosling\\_shows\\_the\\_best\\_stats\\_you\\_ve\\_ever\\_seen.html](http://www.ted.com/index.php/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen.html). The <http://TED.com> Web page is featured in Time, April 27, 2009, p. 44. In another presentation, Rosling shows an animation of 200 years of health and wealth data of 200 countries in just four minutes: <http://www.flixxy.com/200-countries-200-years-4-minutes>.

htm. The underlying software is accessible at <http://www.gapminder.org/> and it is further described in Rosling & Johansson (2009). Google has acquired the software, now called Google Public Data Explorer (<http://www.google.com/publicdata/home>), and they are planning to further extend it. We will revisit this software once we have discussed interactive and dynamic graphics.

## 1.2 Why Graphics !?!

Why do we need graphics at all. Aren't summary statistics sufficient?

Start R and load the Anscombe (1973) data set. Just type `anscombe` to check whether these data are available — if not, you may have to load the data via:

```
require(stats)
data(anscombe)
```

Then calculate some summary statistics (separately for the four columns of X's and Y's): mean of the X's, mean of the Y's, standard deviation of the X's, standard deviation of the Y's, correlation coefficient, slope and intercept of the regression line, rms error.

```
http://www.math.usu.edu/~symanzik/teaching/2011\_stat6560/RDataAndScripts/Anscombe.R
```

So, the four pairs of X/Y columns basically are identical !?!

But, didn't we forget to **plot** the data !!!

```
http://www.math.usu.edu/~symanzik/teaching/2011\_stat6560/RDataAndScripts/Anscombe2.R
```

See here for additional references:

```
http://en.wikipedia.org/wiki/Anscombe's\_quartet
```

```
http://pbil.univ-lyon1.fr/library/base/html/anscombe.html
```

Tufte (1983), p. 13, concludes:

“Graphics *reveal* data. Indeed graphics can be more precise and revealing than conventional statistical computations. Consider Anscombe's quartet: all four of these data sets are described by exactly the same linear model (at least until the residuals are examined).”

## 1.3 How to Display Data Badly

Wainer (1997), p. 12, states:

**“The aim of good data graphics is to display data accurately and clearly.**

[...]

Thus, if we wish to display data badly, we have three avenues to follow.

A. Don’t show much data.

B. Show the data inaccurately.

C. Obfuscate the data.”<sup>†</sup>

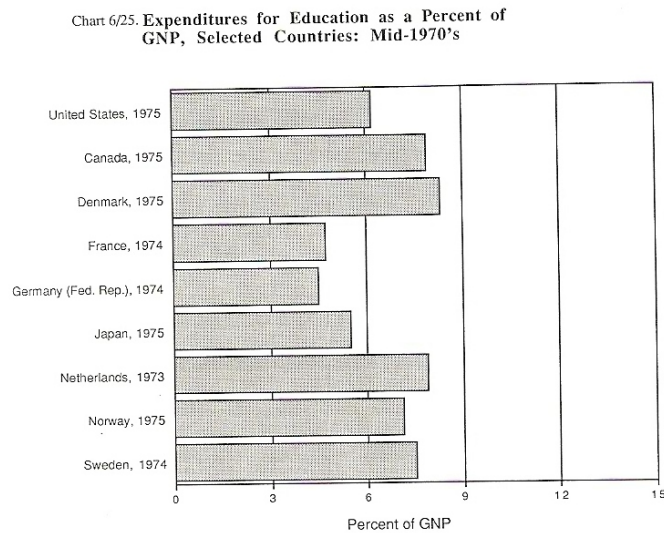
Let us follow these strategies:

---

<sup>†</sup>Show the data unclearly.

### 1.3.1 Don't show much data

**Rule 1: Show as little data as possible (minimize the data density).**



Data Density = 9 numbers / 63 sq. ins. = .14

FIGURE 2. Chart 6/25 from *Social Indicators III* showing expenditures for education for nine countries as a function of GNP.

Figure 2: Wainer (1997), p. 13, Figure 2.

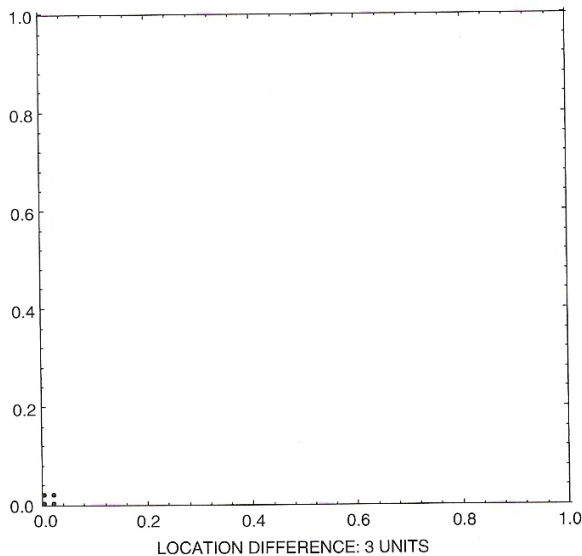


FIGURE 3. A graph of obviously low data density.

Figure 3: Wainer (1997), p. 13, Figure 3.

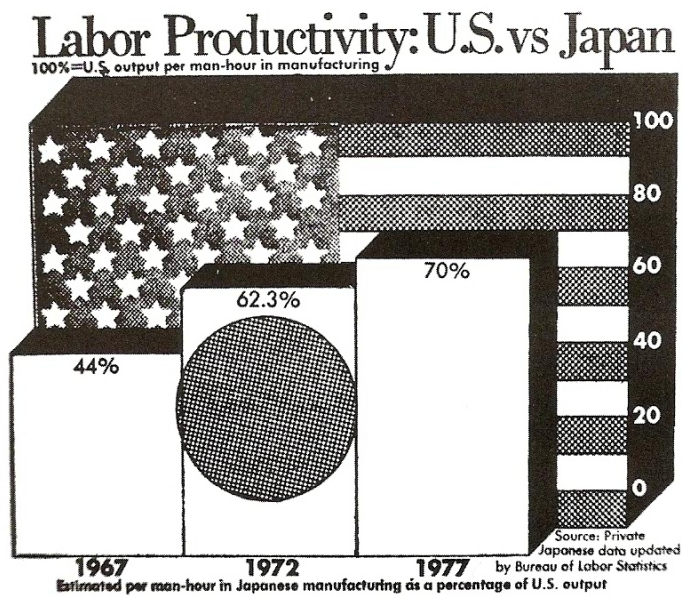


FIGURE 6. A graph with low data density filled in with chartjunk from the *Washington Post*, 1978.

Figure 4: Wainer (1997), p. 16, Figure 6.

Rule 2: Hide what data you do show (minimize the data/ink ratio).

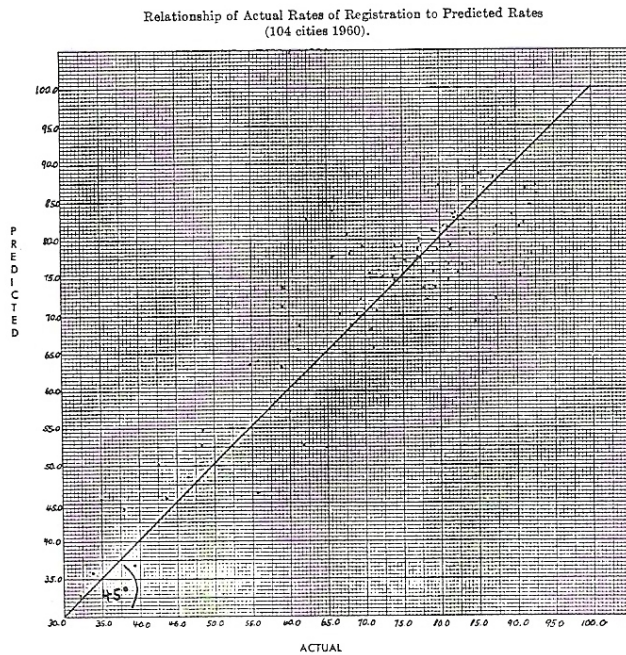


FIGURE 8. Hiding the data in the grid.

Figure 5: Wainer (1997), p. 17, Figure 8: Hiding the data in the grid.

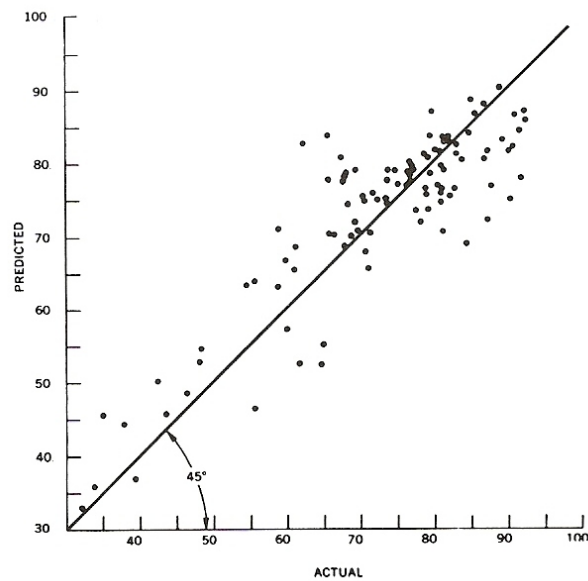
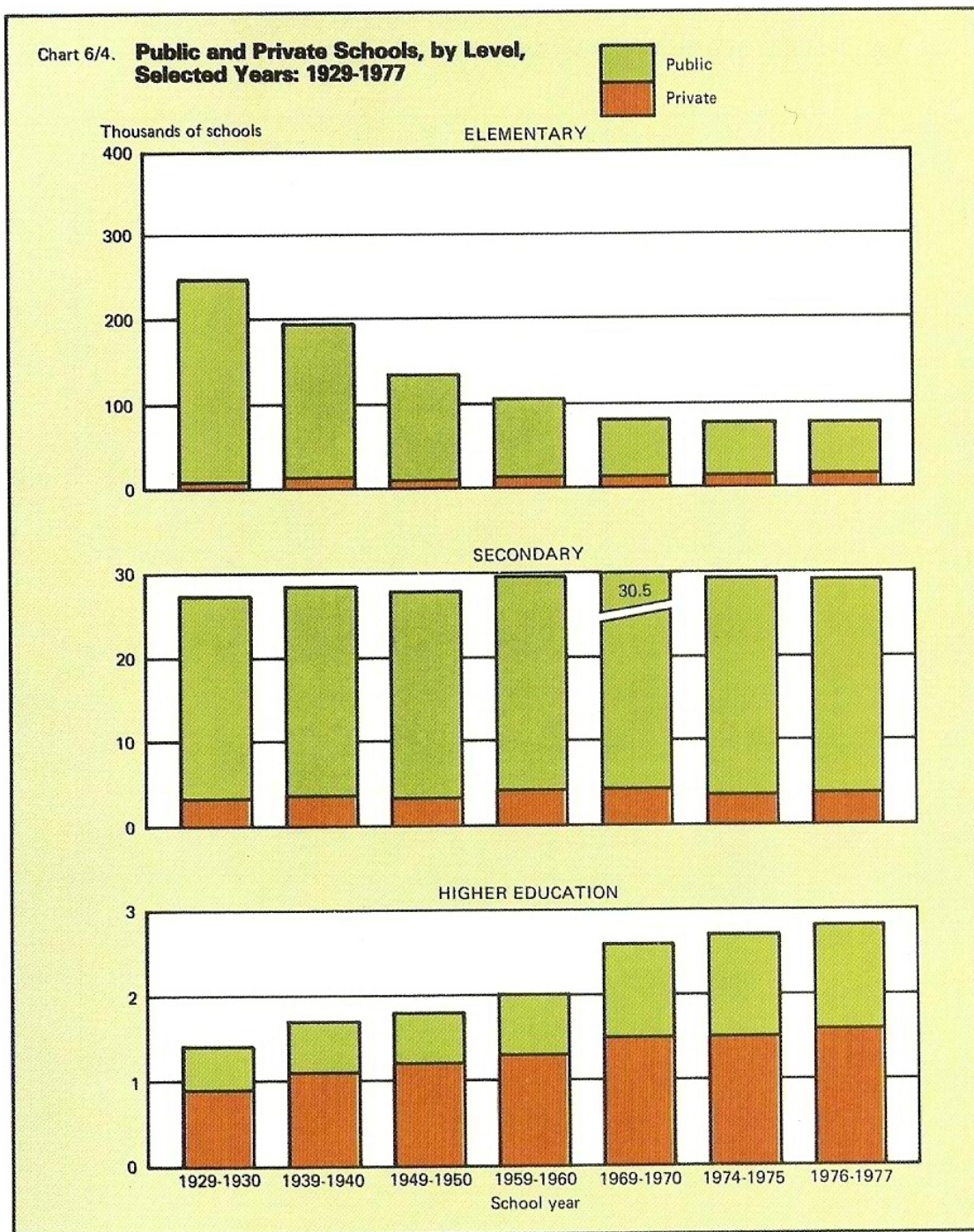


FIGURE 10. A redone example of the data from figure 8.

Figure 6: Wainer (1997), p. 18, Figure 10: Wainer (1997), p. 17, Figure 8, improved.





CHAPTER 1, FIGURE 11. Hiding the data in the scale.

Figure 7: Wainer (1997), p. 20A, Figure 11: Hiding the data in the scale.

FIGURE 12. Expanding the scale and showing the data for the number of private elementary schools from figure 11.

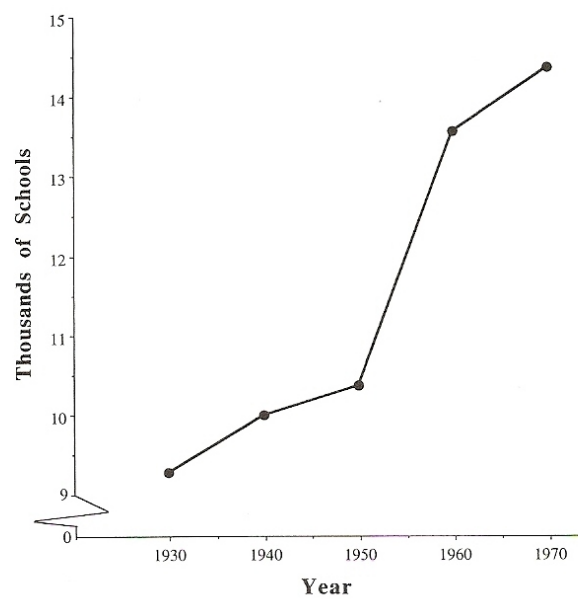


Figure 8: Wainer (1997), p. 20, Figure 12: Wainer (1997), p. 20A, Figure 11, improved.



### 1.3.2 Show the data inaccurately

Rule 3: Ignore the visual metaphor altogether.

FIGURE 13. Ignoring the visual metaphor by letting a longer bar segment represent a smaller amount of coal (from the *New York Times*, 1978).

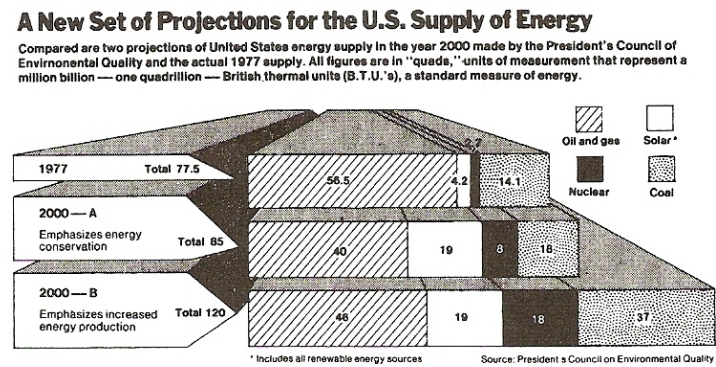


Figure 9: Wainer (1997), p. 20, Figure 13.

### U.S. trade with China and Taiwan

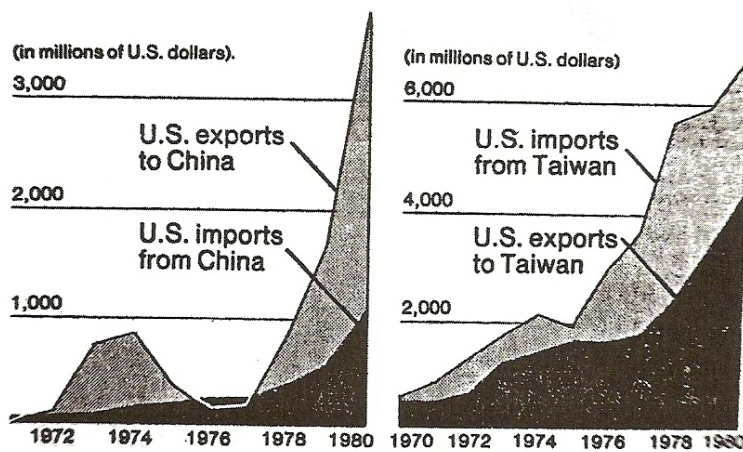


FIGURE 14. Reversing the metaphor in mid-graph while changing scales on both axes (from the *New York Times*, June 14, 1981).

Figure 10: Wainer (1997), p. 21, Figure 14.

FIGURE 15. Figure 14 redone with a consistent scale and visual metaphor.

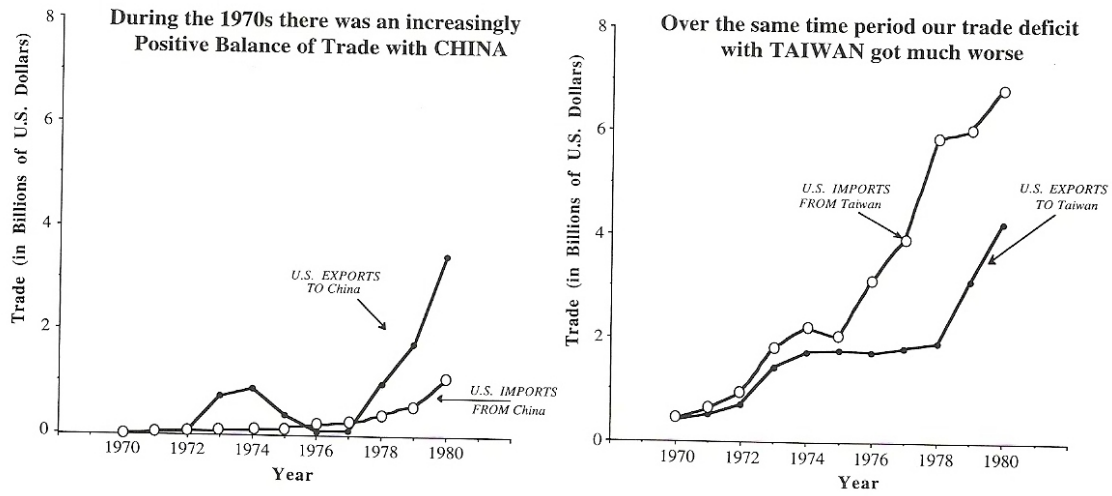


Figure 11: Wainer (1997), p. 21, Figure 15: Wainer (1997), p. 21, Figure 14, improved.

Rule 4: Only order matters.

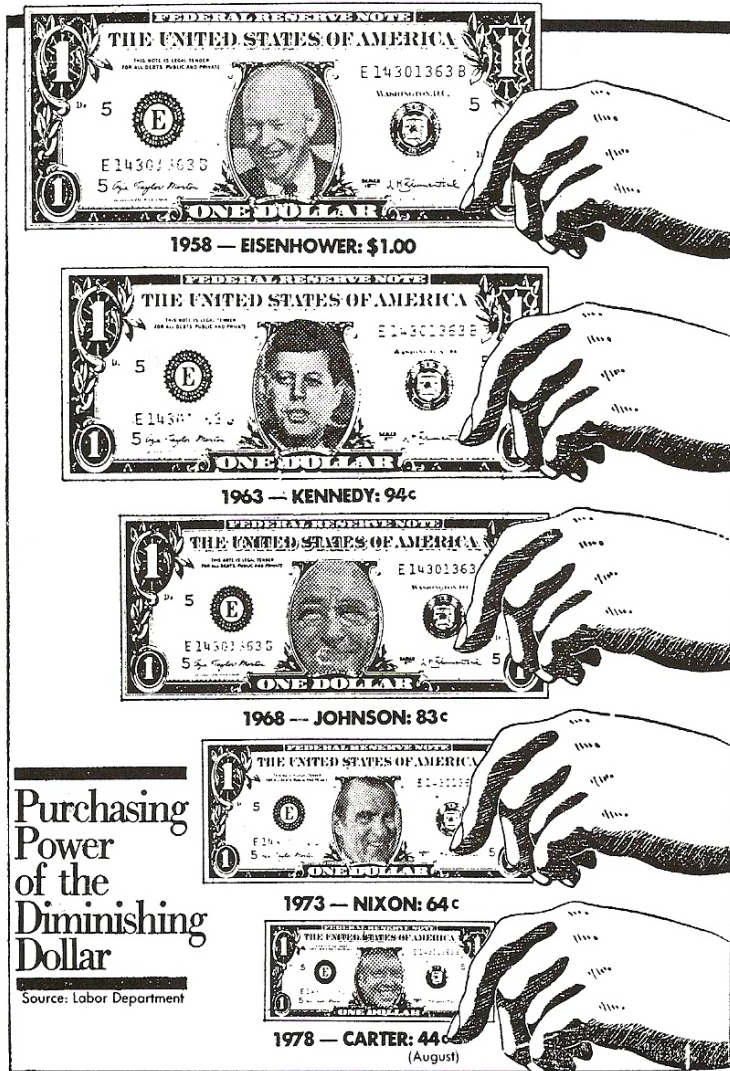


FIGURE 17. An example of how to goose up the effect by squaring the eyeball.

Figure 12: Wainer (1997), p. 23, Figure 17.

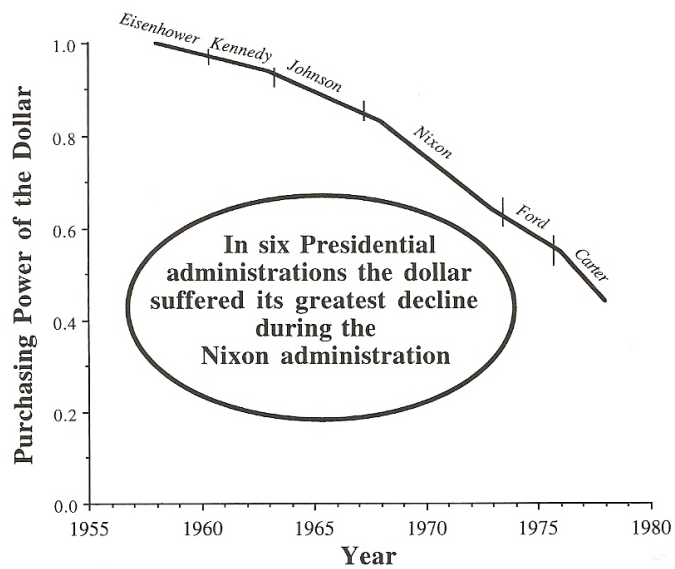


FIGURE 18. The data in figure 17 as an unadorned line chart (from Wainer, 1980).

Figure 13: Wainer (1997), p. 24, Figure 18: Wainer (1997), p. 23, Figure 17, improved.

FIGURE 19. Cubing the visual effect and choosing the origin to yield a near record lie factor of over 131,000% (from the *Washington Post*).

### U.S. Beer Sales and Schlitz's Share

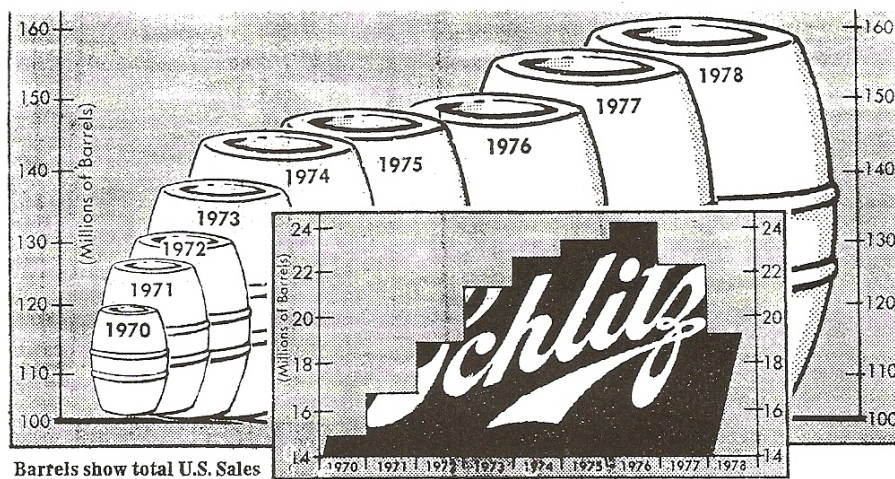


Figure 14: Wainer (1997), p. 24, Figure 19.

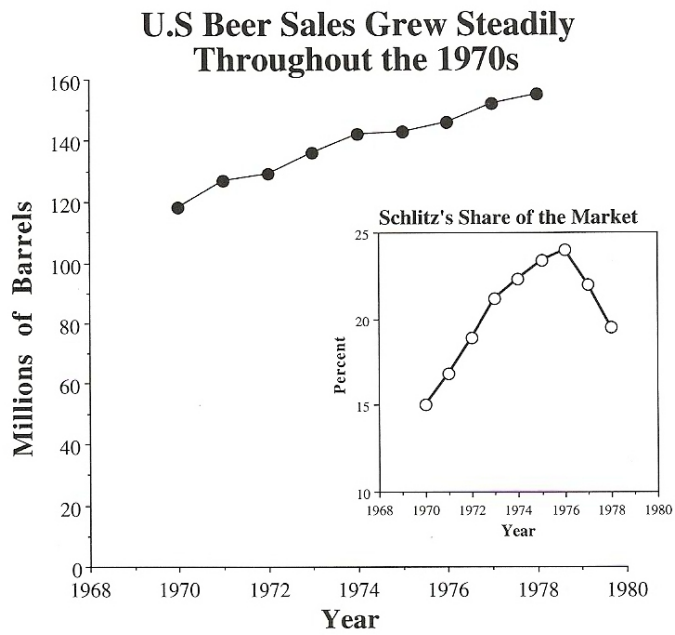


FIGURE 20. Data from figure 19 redone without tricks (from Wainer, 1980).

Figure 15: Wainer (1997), p. 25, Figure 20: Wainer (1997), p. 24, Figure 19, improved.



Rule 5: Graph data out of context.

FIGURE 21. Hiding the effect by the careful choice of scale and origin (from the *Washington Post*).

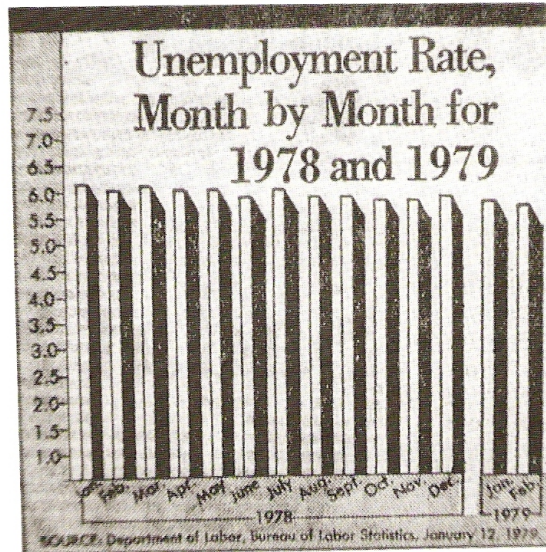


Figure 16: Wainer (1997), p. 26, Figure 21.

FIGURE 22. Regraph of data from figure 21 with expanded scale, different starting point, and previous year's average added for context (from Wainer, 1980).



Figure 17: Wainer (1997), p. 26, Figure 22: Wainer (1997), p. 26, Figure 21, improved.

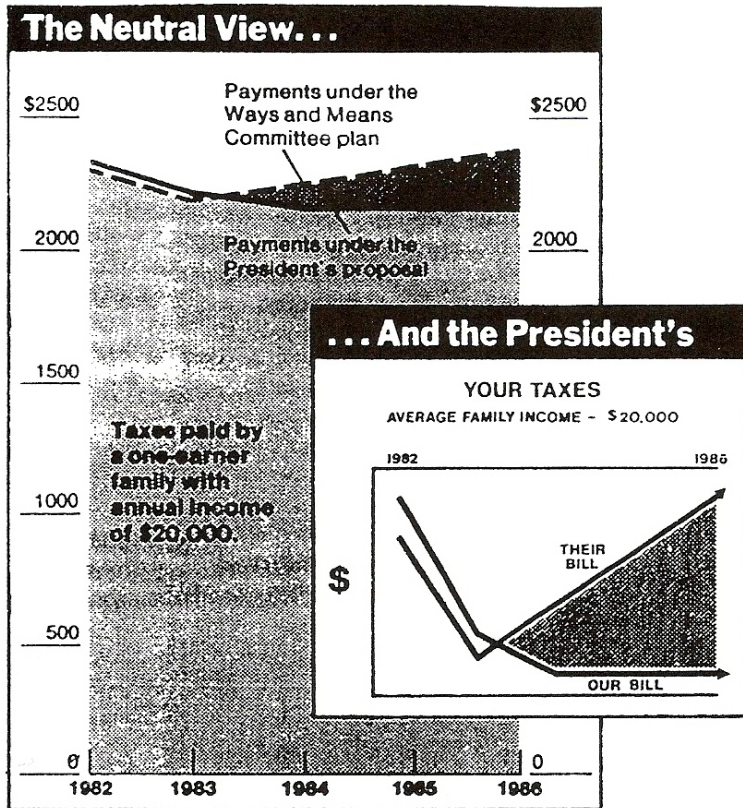


FIGURE 23. *New York Times* graphs showing how lack of context changes our perceptions about alternative tax bills.

Figure 18: Wainer (1997), p. 27, Figure 23.

### 1.3.3 Obfuscate the data

Rule 6: Change scales in mid-axis.

FIGURE 24. Changing the scale in mid-axis to make large differences seem small (from the *New York Post*, May 12, 1981).

## The soaraway Post — the daily paper New Yorkers trust

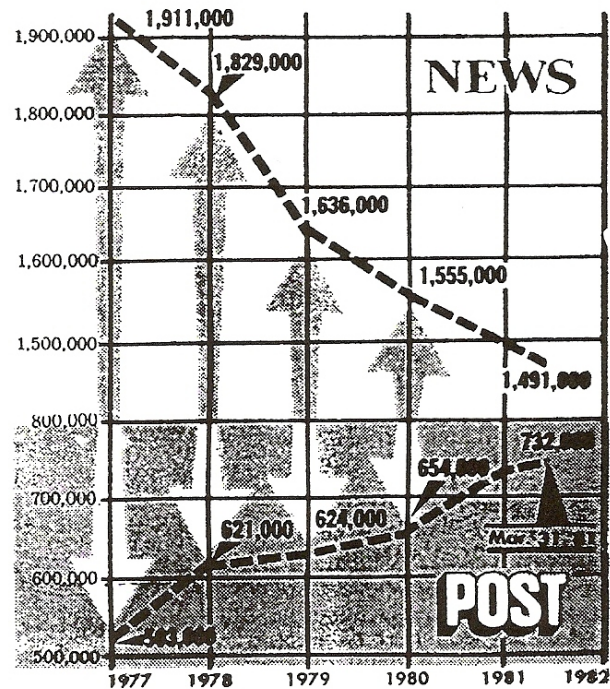


Figure 19: Wainer (1997), p. 28, Figure 24.



FIGURE 25. Changing scale in mid-axis to make exponential growth linear (from the *Washington Post*, Jan. 11, 1979, in an article titled "Pay, Practices of Doctors on Examining Table" by Victor Cohn and Peter Milius).

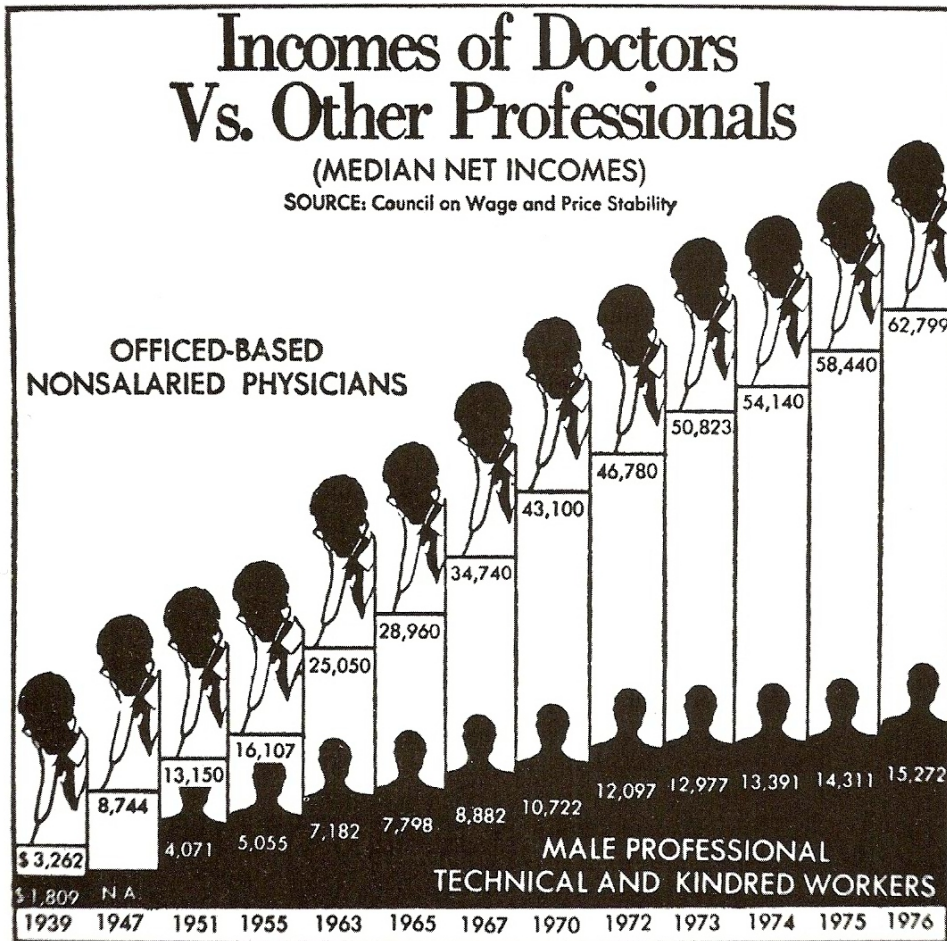


Figure 20: Wainer (1997), p. 29, Figure 25.

FIGURE 26. Data from figure 25 redone with a linear scale (from Wainer, 1980).

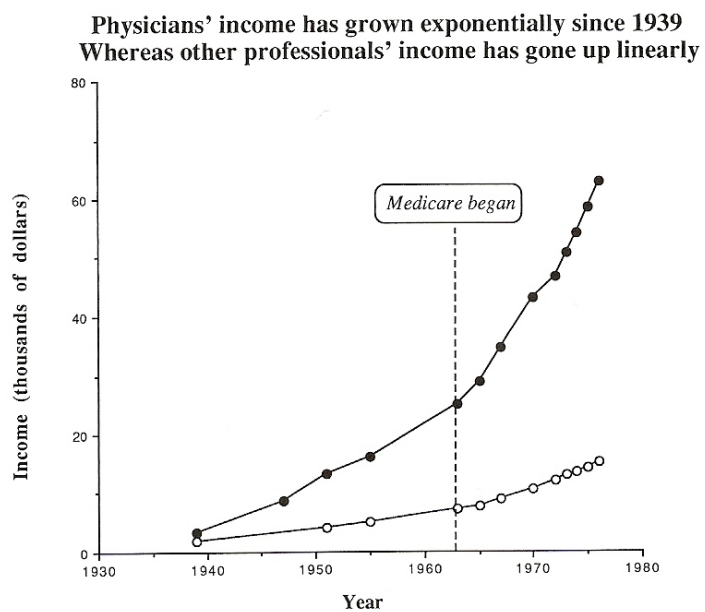
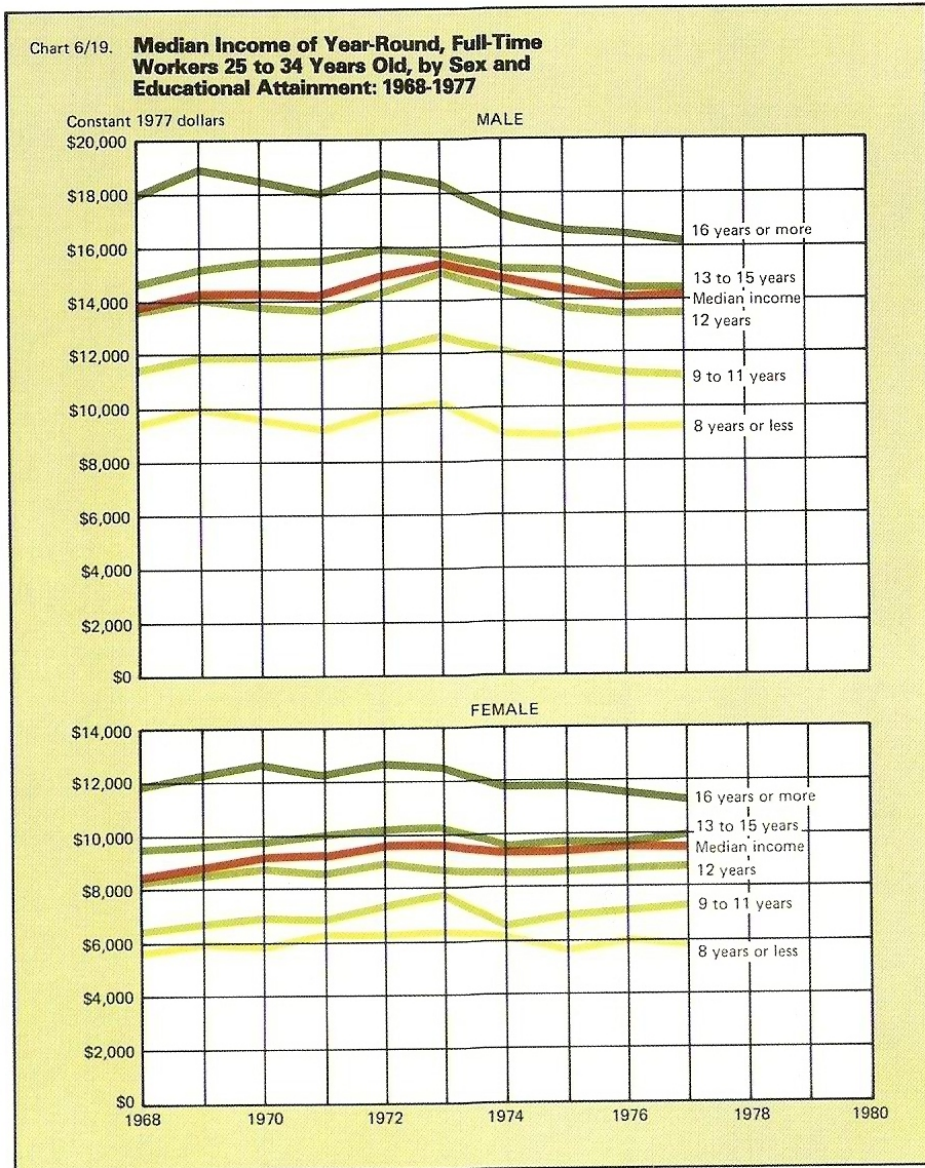


Figure 21: Wainer (1997), p. 30, Figure 26: Wainer (1997), p. 29, Figure 25, improved.

Rule 7: Emphasize the trivial (ignore the important).



CHAPTER 1, FIGURE 27. Emphasizing the trivial: Hiding the main effect of sex differences in income through the vertical placement of plots.

Figure 22: Wainer (1997), p. 20A, Figure 27.

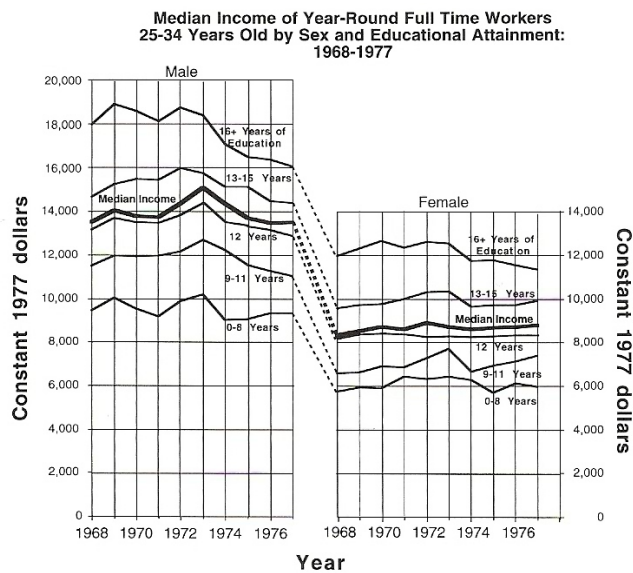


FIGURE 28. Figure 27 redone with the two plots horizontally opposed, showing the size of sex differences more clearly.

Figure 23: Wainer (1997), p. 31, Figure 28: Wainer (1997), p. 20A, Figure 27, improved.

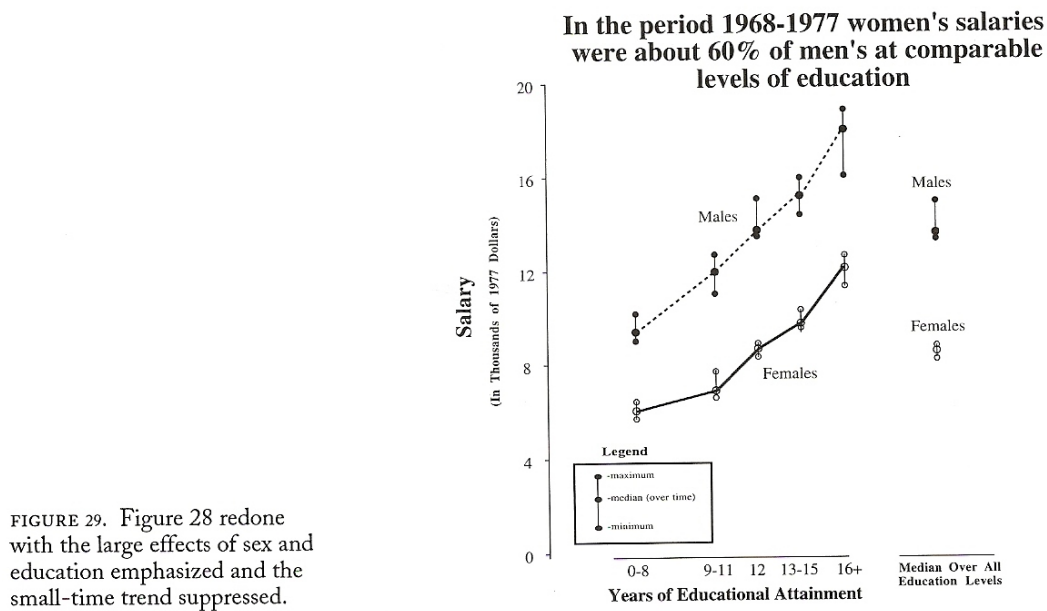
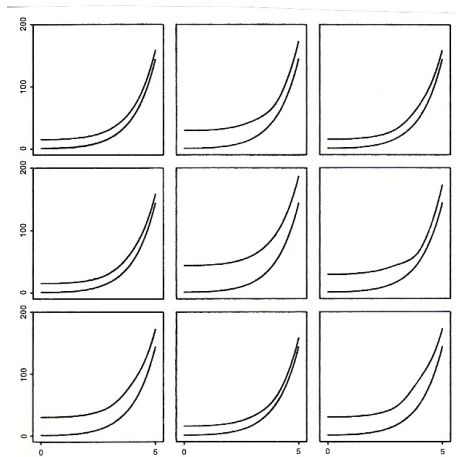


FIGURE 29. Figure 28 redone with the large effects of sex and education emphasized and the small-time trend suppressed.

Figure 24: Wainer (1997), p. 32, Figure 29: Wainer (1997), p. 31, Figure 28, further improved.

## Rule 8: Jiggle the baseline.

FIGURE 30. A graphical experiment (from Cleveland and McGill, 1984). Without looking at the corresponding right panel, try to determine the difference between the two curves in the left panel.



Sorry, these plots  
got scrambled...

Figure 25: Wainer (1997), p. 33, Figure 30.

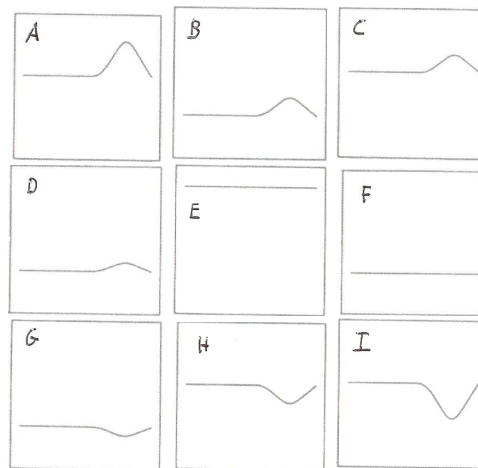
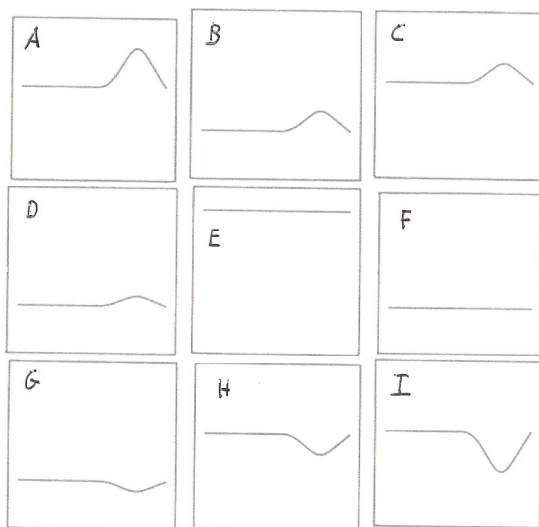
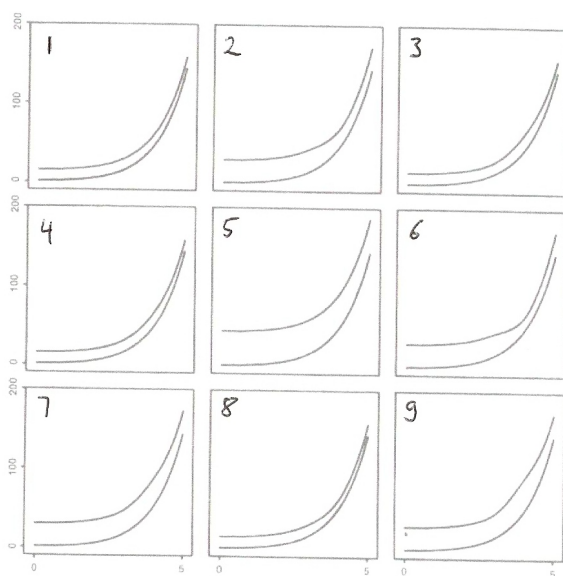


Figure 26: Wainer (1997), p. 33, Figure 30: Scrambled differences. The horizontal axis covers the interval 0 to 5, the vertical axis covers the interval 0 to 50.

# Worksheet

Your Name: \_\_\_\_\_



**Task:** Match each original (labeled 1 to 9) with the plot (labeled A to I) that shows the difference between upper and lower line in the original plot.

**Answer:**

1: \_\_\_\_\_ 2: \_\_\_\_\_ 3: \_\_\_\_\_  
 4: \_\_\_\_\_ 5: \_\_\_\_\_ 6: \_\_\_\_\_  
 7: \_\_\_\_\_ 8: \_\_\_\_\_ 9: \_\_\_\_\_



FIGURE 31. William Playfair's eighteenth-century graph of England's imports and exports with the East Indies (from Cleveland and McGill, 1984).

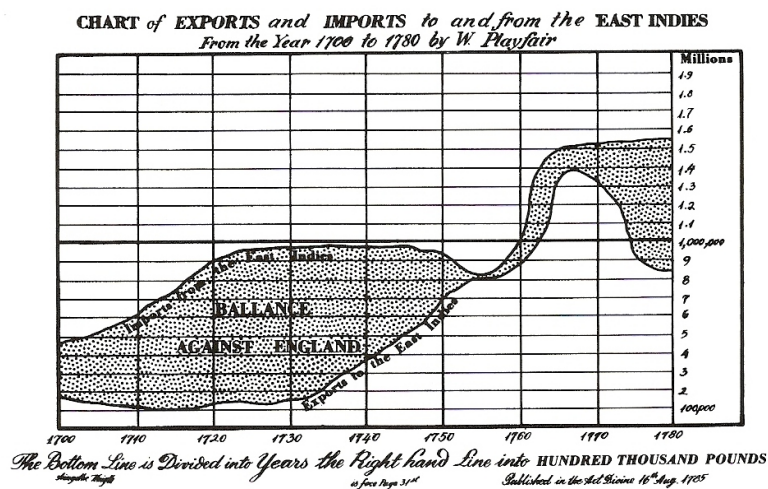


Figure 27: Wainer (1997), p. 34, Figure 31: One of William Playfair's few mistakes.

FIGURE 32. A graph of the difference between West Indies imports and exports showing explicitly the previously invisible jump in the 1760s (from Cleveland and McGill, 1984).

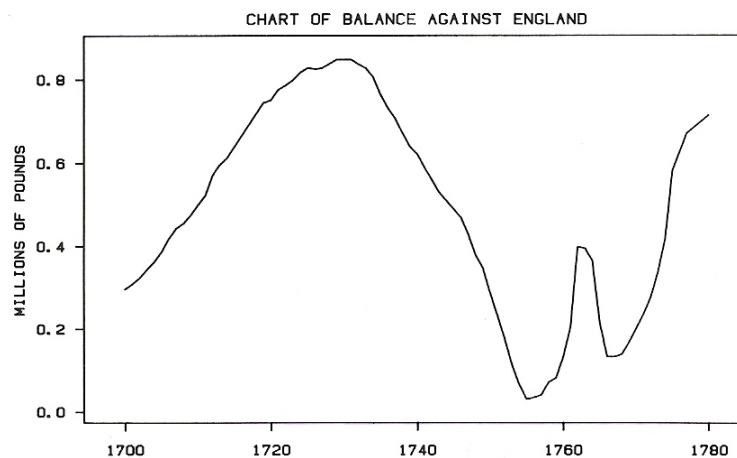


Figure 28: Wainer (1997), p. 34, Figure 32: Wainer (1997), p. 34, Figure 31, improved.

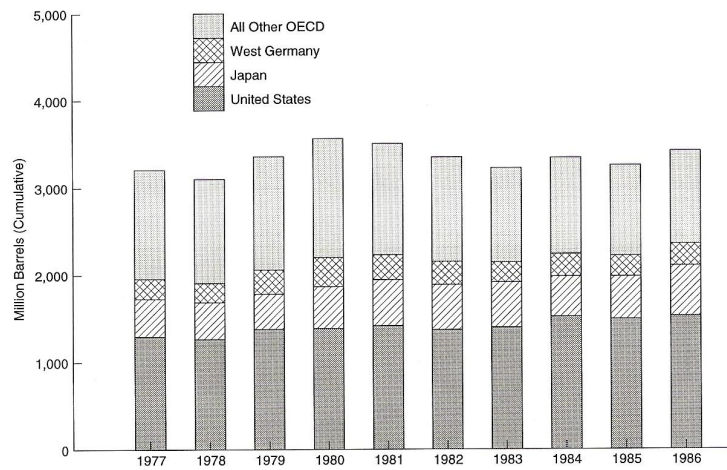


FIGURE 33. From the U.S. Department of Energy's *Annual Energy Review, 1986*, showing the changes in primary stocks of petroleum in OECD countries.

Figure 29: Wainer (1997), p. 35, Figure 33.

OECD PETROLEUM STOCKS HAVE STABILIZED  
But Not All Countries Are Pulling Their Own Weight

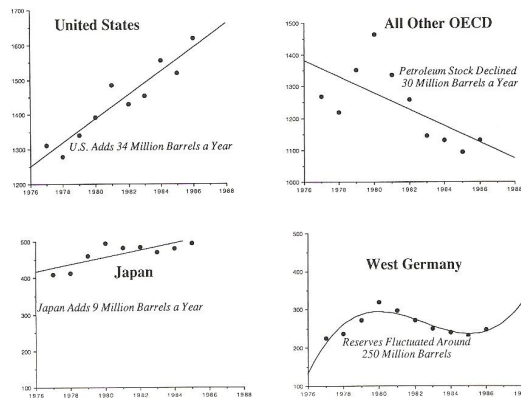
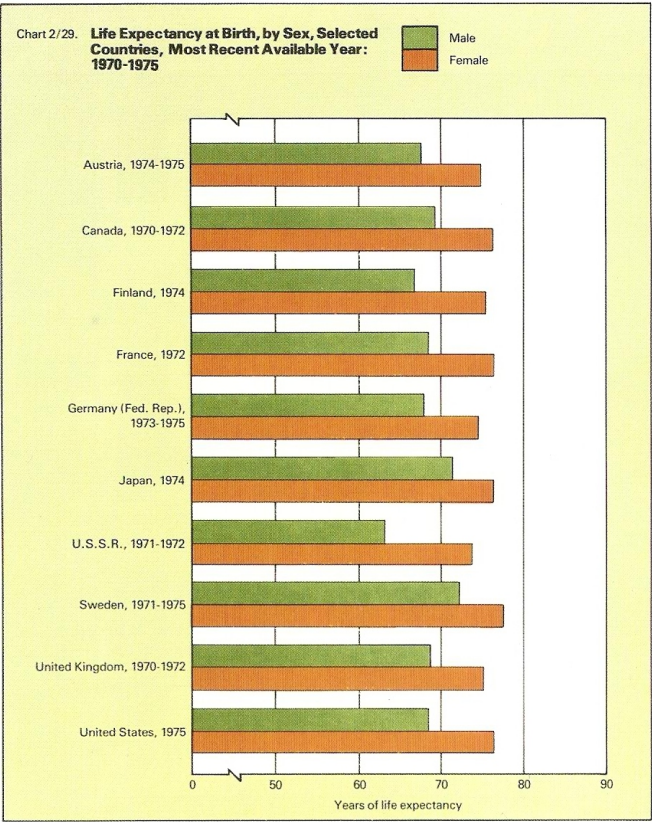


FIGURE 34. Regraphing of the data from figure 33 in which each country's data are shown relative to a straight line.

Figure 30: Wainer (1997), p. 36, Figure 34: Wainer (1997), p. 35, Figure 33, improved.



Rule 9: Alabama first!



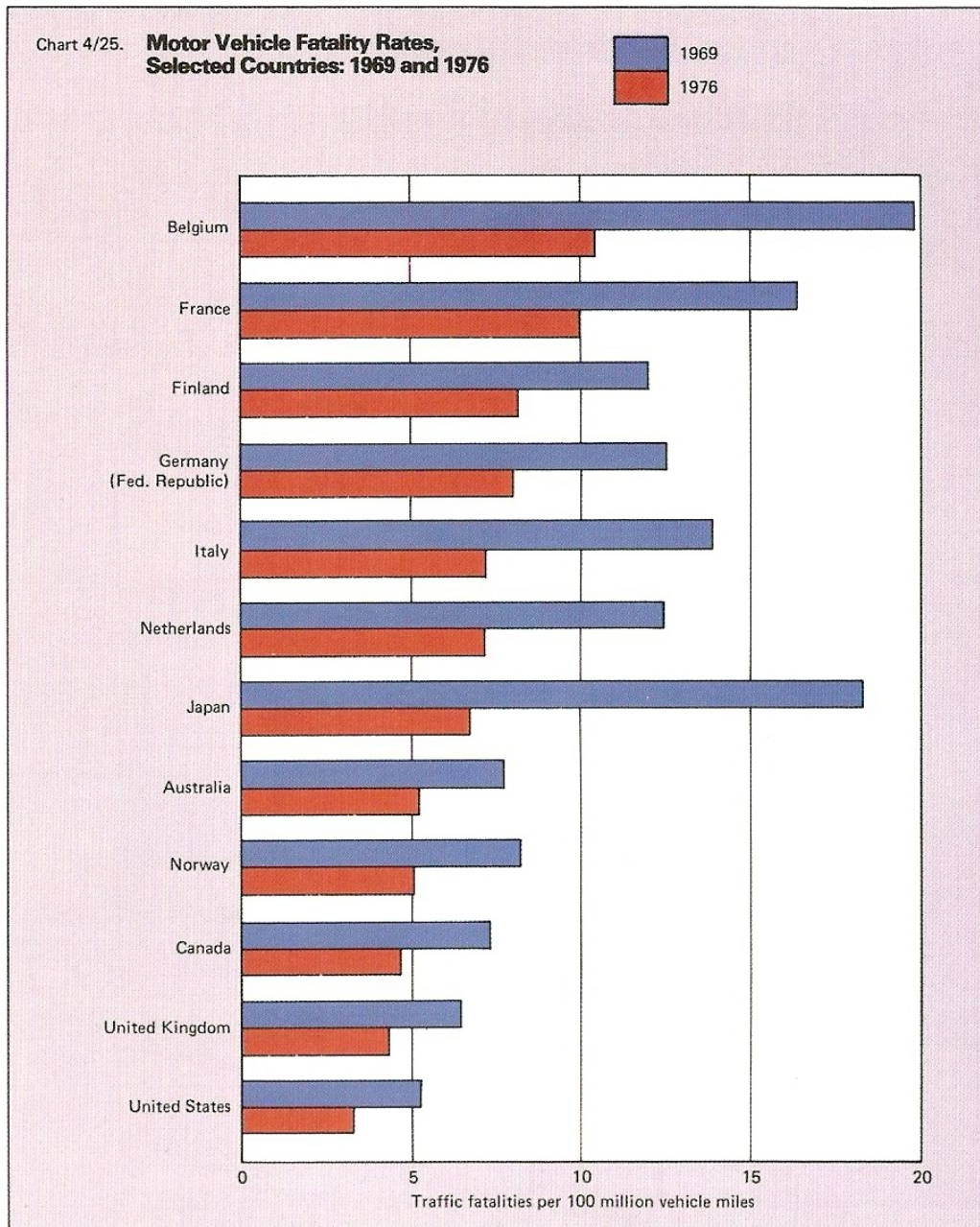
CHAPTER 1, FIGURE 35. Austria first! Obscuring the data structure in some life expectancy data by alphabetizing the plot.

Figure 31: Wainer (1997), p. 20B, Figure 35.



FIGURE 36. Ordering and spacing the data from figure 35 as a stem-and-leaf diagram provides insights previously invisible.

Figure 32: Wainer (1997), p. 37, Figure 36: Wainer (1997), p. 20B, Figure 35, improved.



CHAPTER 1, FIGURE 37. Ordering the bar chart by the data tells the tale a bit more clearly.

Figure 33: Wainer (1997), p. 20B, Figure 37: Layout similar to Wainer (1997), p. 20B, Figure 35, but improved due to ordering.

Rule 10: Label: (a) illegibly, (b) incompletely, (c) incorrectly, and (d) ambiguously.

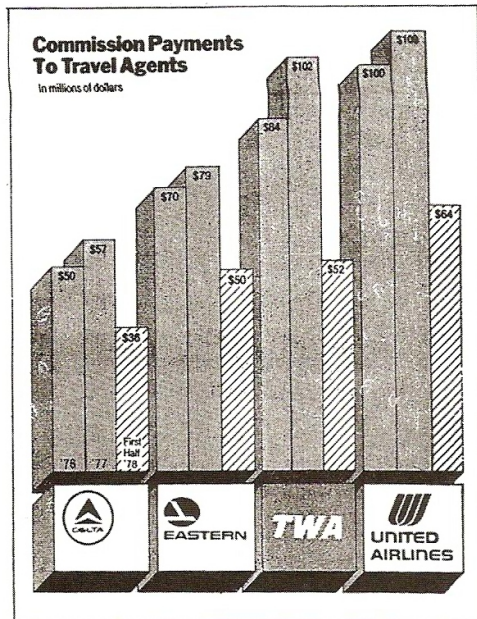


FIGURE 38. Mixing a changed metaphor with a tiny label reverses the meaning of the data.

Figure 34: Wainer (1997), p. 39, Figure 38.

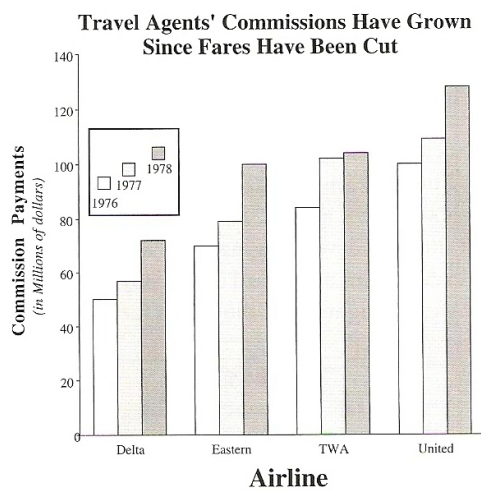


FIGURE 39. Figure 38 redrawn with 1978 data placed on a comparable basis shows that the fare cuts have been a boon to travel agents.

Figure 35: Wainer (1997), p. 39, Figure 39: Wainer (1997), p. 39, Figure 38, improved.

Rule 11: More is murkier: (a) more decimal places and (b) more dimensions.

<b>TABLE 1</b>		
<b>Life Expectancy at Birth</b>		
<b>Country</b>	<b>Male</b>	<b>Female</b>
Argentina	56.90	61.40
Brazil	39.30	45.50
Canada	67.61	72.92
Iceland	66.10	70.30
Japan	65.37	70.26
Mexico	37.92	39.79
Netherlands	71.40	74.80
New Zealand	68.20	73.00
Norway	71.11	74.70
Spain	58.76	63.50

Figure 36: Wainer (1997), p. 40, Table 1.

<b>TABLE 2</b>		
<b>Life Expectancy at Birth</b>		
<b>Country</b>	<b>Male</b>	<b>Female</b>
Netherlands	71	75
Norway	71	75
New Zealand	68	73
Canada	68	73
Iceland	66	70
Japan	65	70
Spain	59	64
Argentina	57	61
Brazil	39	46
Mexico	38	40

Figure 37: Wainer (1997), p. 40, Table 2: Wainer (1997), p. 40, Table 1, improved.



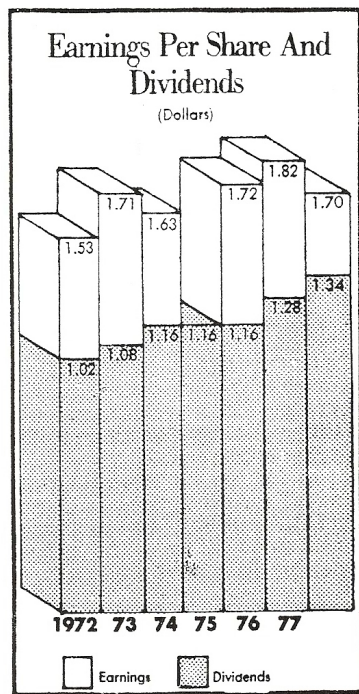


FIGURE 41. An extra dimension on earnings and dividends confuses even the grapher (from the *Washington Post*, 1979).

Figure 38: Wainer (1997), p. 41, Figure 41.

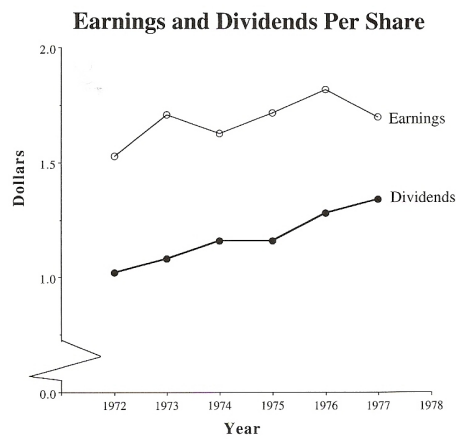
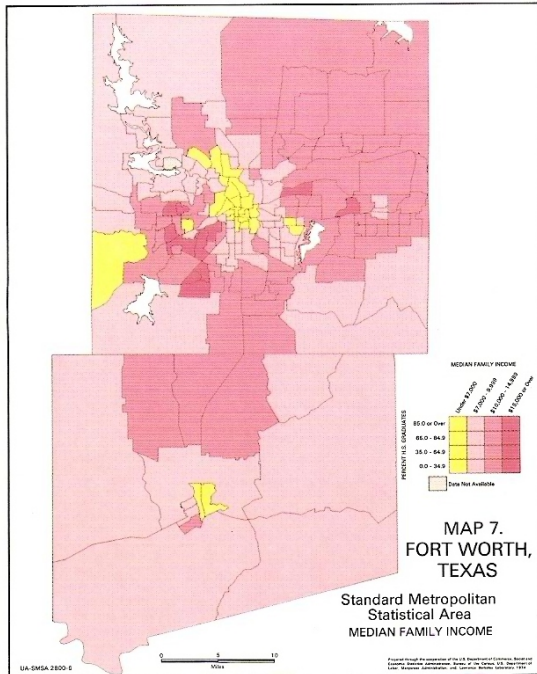


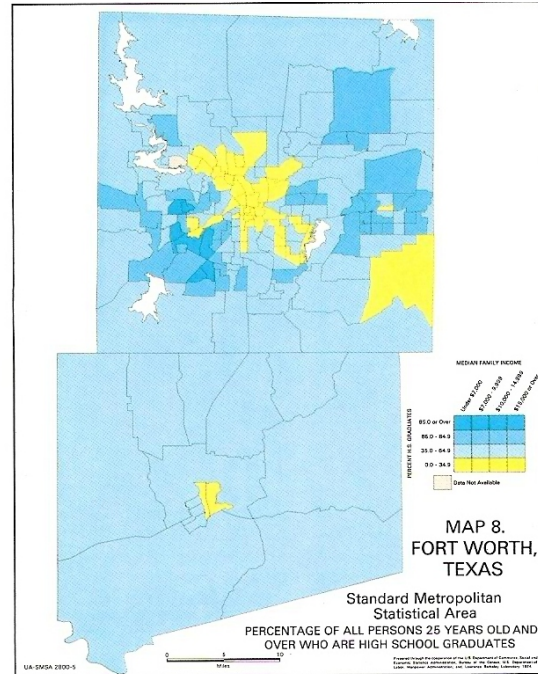
FIGURE 42. Data from figure 41 redrawn simply.

Figure 39: Wainer (1997), p. 42, Figure 42: Wainer (1997), p. 41, Figure 41, improved.

Rule 12: If it has been done well in the past, think of a new way to do it.

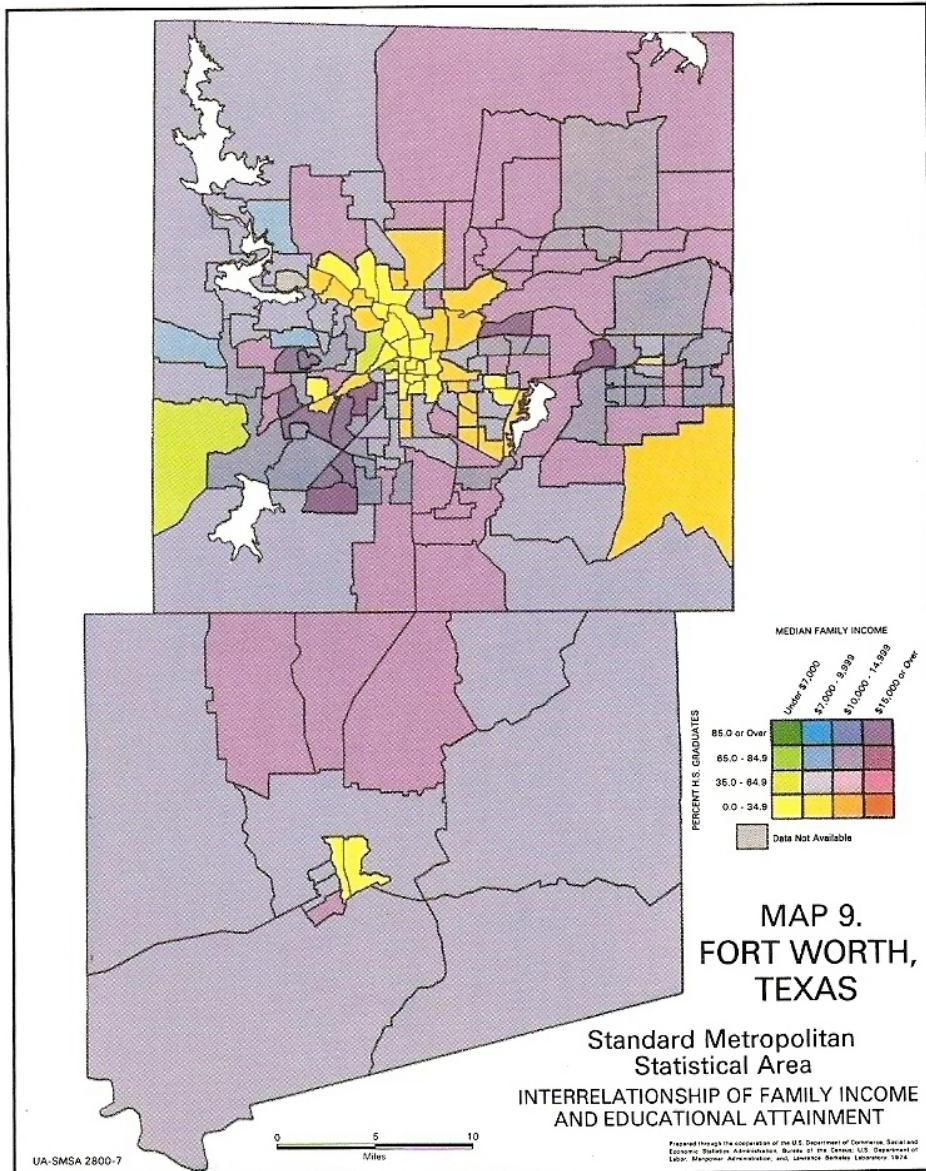


CHAPTER 1, FIGURE 44. The geographic distribution of median family income in Fort Worth, Texas, in 1974.



CHAPTER 1, FIGURE 45. The geographic distribution of percentage of high-school graduates in Fort Worth, Texas, in 1974.

Figure 40: Wainer (1997), p. 20C, Figures 44 & 45: Traditionl maps.



CHAPTER 1, FIGURE 46. The geographic distribution of both median family income and percentage of high-school graduates in Fort Worth, Texas, in 1974, shown as a two-variable color map.

Figure 41: Wainer (1997), p. 20C, Figure 46: Wainer (1997), p. 20C, Figures 44 & 45, modified but **not** improved.



## 1.4 Bad Graphics are Everywhere — In Space and in Time

Example 1: Zion National Park, UT, Shuttle Parking Lot, December 28, 2002

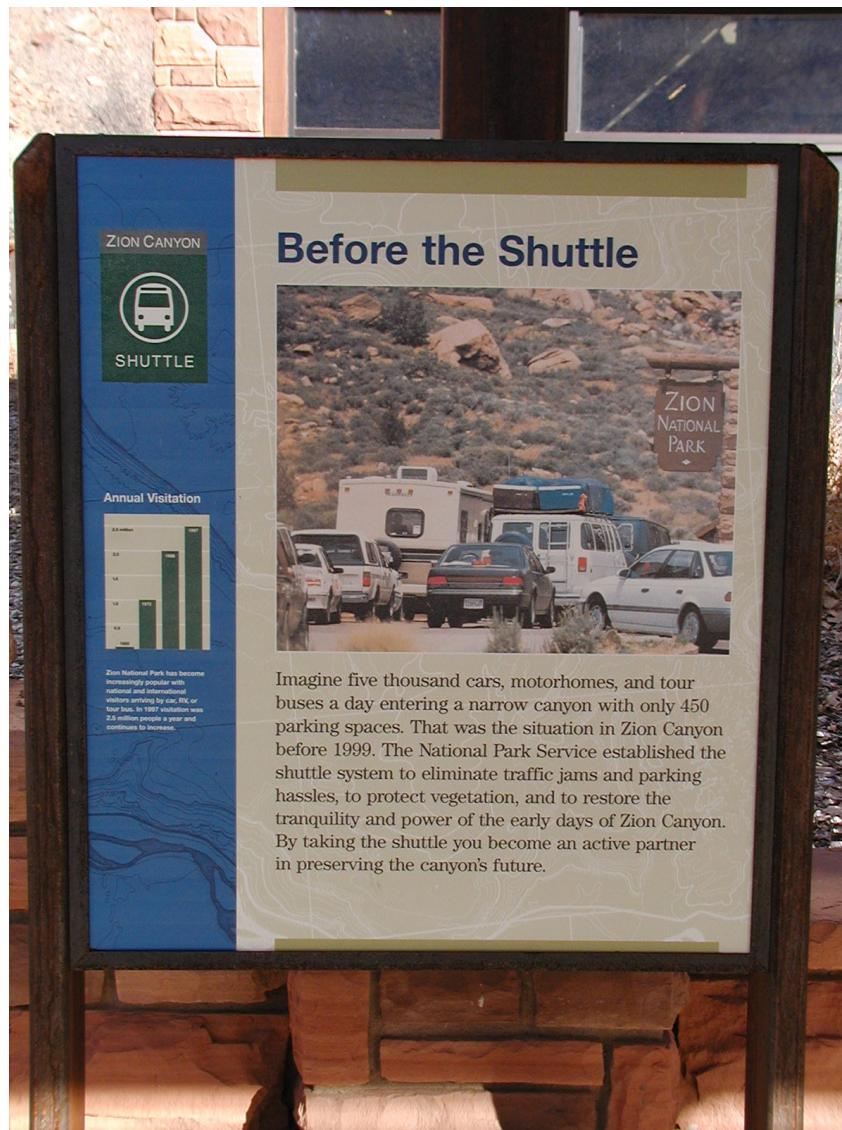


Figure 42: Personal Photograph: From the distance, the annual visitation appears to increase linearly, . . .



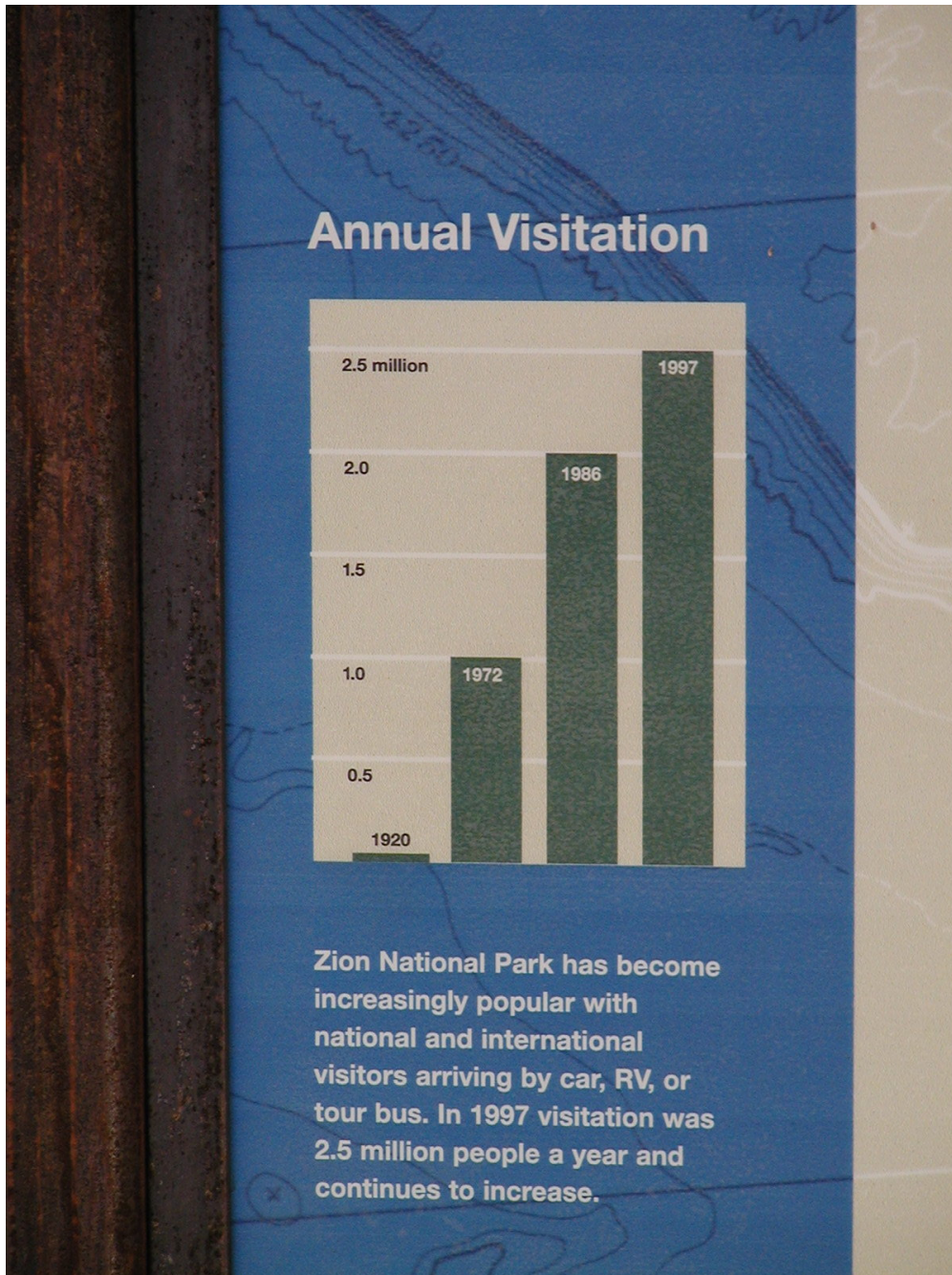


Figure 43: Personal Photograph: ... but at a closer view, this is certainly not the case.

### Rules followed (to make this a bad graphic):

- Rule 6: Change scales in mid-axis.

Years on the horizontal axis are 1920, 1972, 1986, and 1997, i.e., the gaps are 52, 14, and 11 years. However, the same spacing has been used.

- Rule 10: Label: (a) illegibly, (b) incompletely, (c) incorrectly, and (d) ambiguously.

No axis label on vertical axis; what are 2.5 millions, etc. — visitors or cars? Also, no label on the horizontal axis. Moreover, listing the year near the top of each of the bars could be confusing as this might be interpreted as the actual number of visitors (in millions or so).

- Rule 1: Show as little data as possible (minimize the data density).

There are only 4 data points, but the figure is considerably filled with ink used for the bars.

- Rule 5: Graph data out of context.

Data are shown for only 4 years. However, the data range is 77 years. Why have these 4 years been chosen — and not any others (and in particular, why not more years)?

### Improved Version:

[http://www.math.usu.edu/~symanzik/teaching/2011\\_stat6560/RDataAndScripts/Zion.R](http://www.math.usu.edu/~symanzik/teaching/2011_stat6560/RDataAndScripts/Zion.R)

Example 2: Berlin, Germany, August 20, 2006



Figure 44: Personal Photograph: Exhibit at the 1936 Berlin Olympic Site, related to the history of the Olympic area from 1909 to 1936 to 2006.





Figure 45: Personal Photograph: Historical graphic (from the late 1920ies), dedicated to the development of women's gymnastics as part of the *Deutsche Turnerschaft* (the governing body of German gymnastics).

### Rules followed (to make this a bad graphic):

- Rule 6: Change scales in mid-axis.

Years on the horizontal axis are 1897, 1900, 1904, 1907, 1914, 1919, 1921, 1924, and 1927, i.e., the gaps are 3, 4, 3, 7, 5, 2, 3, and 3 years. However, the same spacing has been used.

- Rule 4: Only order matters.

The size of the figures is not proportional to the numbers presented. Moreover, which part of the figure represents a value? (Look at the raised arms in 1904 and 1924.)

- Rule 3: Ignore the visual metaphor altogether.

The figure for 1919 is smaller than that for 1897, 1900, 1904, and 1907 — although the value is bigger for 1919. Moreover, two different scales are used for the vertical axis in the upper and in the lower part of the figure.

### Improved Version:

[http://www.math.usu.edu/~symanzik/teaching/2011\\_stat6560/RDataAndScripts/Berlin.R](http://www.math.usu.edu/~symanzik/teaching/2011_stat6560/RDataAndScripts/Berlin.R)

### Example 3: Wikipedia, 2009

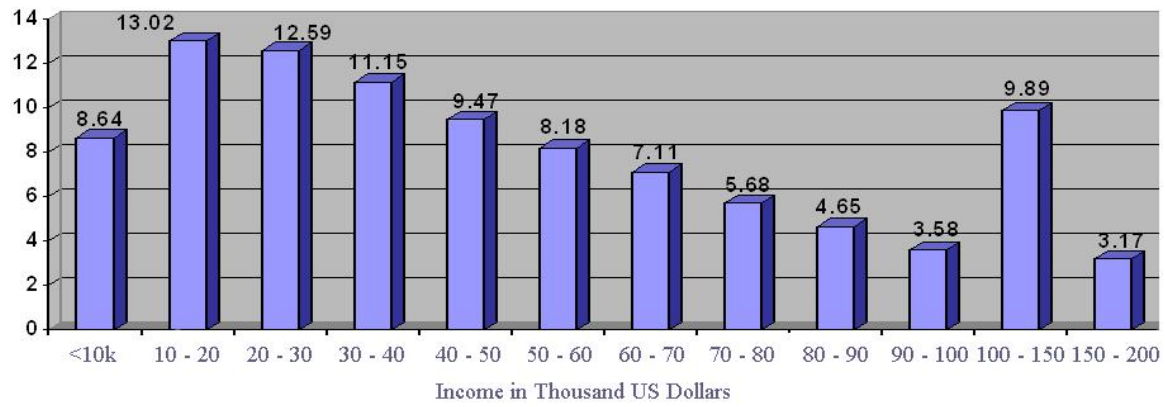


Figure 46: Figure taken from [http://en.wikipedia.org/wiki/Household\\_income\\_in\\_the\\_United\\_States](http://en.wikipedia.org/wiki/Household_income_in_the_United_States) on 1/13/2009.

### Rules followed (to make this a bad graphic):

- Incorrect plot type!!!

Income is quantitative and not categorical. We need a histogram here and not a bar chart.

- Rule 3: Ignore the visual metaphor altogether.

or: Rule 6: Change scales in mid-axis.

The 100–150 Thousand Dollars income interval seems to be the most outstanding interval, but this is due to the fact that this is a 50 Thousand Dollars wide interval. Most other intervals are only 10 Thousand Dollars wide. Histograms need to be drawn using the density scale if class intervals are differently wide, i.e., percentages have to be recalculated as percentage per unit.

- Rule 11: More is murkier: (a) more decimal places and (b) more dimensions.

No need for a third dimension for the bars. Also, no need to list two decimals for the percentages (e.g., 13.02).

- Rule 5: Graph data out of context.

Only incomes up to 200 Thousand Dollar are shown. But 2.87% of the incomes (not shown) are above 200 Thousand Dollars. Not showing these high incomes (and not even mentioning these incomes) is quite misleading.

- Rule 10: Label: (a) illegibly, (b) incompletely, (c) incorrectly, and (d) ambiguously.

Which year is this? The Wikipedia Web page deals with income data from 2004, 2005, and 2006 — so it is not clear which year is the basis for the data used in this figure.

### Improved Version:

[http://www.math.usu.edu/~symanzik/teaching/2011\\_stat6560/RDataAndScripts/Wiki.R](http://www.math.usu.edu/~symanzik/teaching/2011_stat6560/RDataAndScripts/Wiki.R)



## Example 4: Computational Statistics, 2002

430

age, surface and contour graphs. Various types of graphs created by KyPlot are shown in Figures 3 and 4.

Almost every component of each graph can be customized through dialog boxes. Double-clicking an axis of a graph brings up a dialog box through which one can change various settings for the axis interactively. The scales of the x- and y-axes of graphs can be individually set as either linear or logarithmic. Error bars can be attached to either x- or y-values, or both, and the attributes of individual error bars can also be customized. (For example, in Figure 3A, the error bars for two data points of a line graph have been partially suppressed to avoid overlapping.) A break along an axis can be set, over a specific range and at a specific location, to indicate that a range of values has been omitted (Figure 3B).

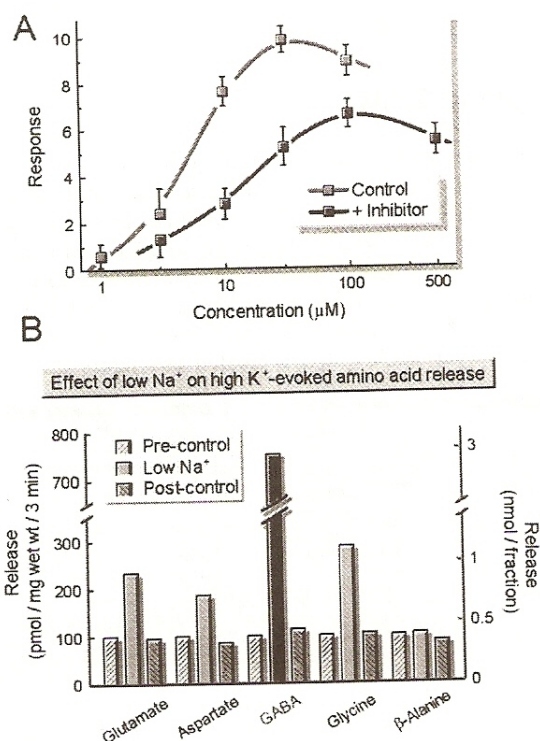


Figure 3: Line and bar graphs created with KyPlot

Figure 47: Yoshioka (2002), p. 430, Figure 3: Intended (!) features of the KyPlot software package for statistical data analysis and visualization.

## Rules followed (to make this a bad graphic):

### 3A:

- Rule 10: Label: (a) illegibly, (b) incompletely, (c) incorrectly, and (d) ambiguously.

Concentration is drawn using a log10-scale. This is not stated (and also not immediately clear with only one additional (unlabeled) ticmark. Moreover, just this additional ticmark (that is not labeled at all!) halfway between two labeled ticmarks makes it very difficult to read off concentration values. Reconstruction of these two missing labels as 3.2 and 32 requires some careful considerations.

- Rule 12: If it has been done well in the past, think of a new way to do it.

People have dealt with overplotting before. We can use different colors (or symbols) for example when parts of the data or information are being overplotted.

- Rule 3: Ignore the visual metaphor altogether.

Except for a concentration of 3.2, the error bars suggest an approximate symmetric (likely normal) distribution of the response. With the error bars partially suppressed, the distribution for the control seems to be skewed to the right and the distribution for the inhibitor seems to be skewed to the left for a concentration of 3.2. Otherwise, with error bars plotted for this concentration level as well, the likely message would be that there is no significant difference between control and inhibitor for a concentration of 3.2 (whereas there is a significant difference for the concentrations of 10, 32, and 100).

### 3B:

- Rule 6: Change scales in mid-axis.

There is a break in the vertical axis. 300 is followed by 700, but the distance between these two values is about the same as for differences of 100 elsewhere on the vertical axis.

- Rule 10: Label: (a) illegibly, (b) incompletely, (c) incorrectly, and (d) ambiguously.

There are two labeled vertical axes! Which of these axes/labels is used, and which is not used?

- Rule 9: Alabama first!

Even worse, there is no sorting at all here.

- Rule 7: Emphasize the trivial (ignore the important).

Five bars, i.e., some considerable amount of space, are used to display that the Pre-control is 100 for each of the five amino acids under investigation. Moreover, it takes a while to realize that the Post-control for all five amino acids is also close to 100 (with some small variation). The response under Low  $\text{Na}^+$  differs considerably, though, for the various amino acids.

- Rule 3: Ignore the visual metaphor altogether.

The “Low  $\text{Na}^+$ ” bar for “GABA” is shaded in black — while all other “Low  $\text{Na}^+$ ” bars are shaded in light gray. We can highlight a single observation (or a subset of observations), but then we should say so in the caption and indicate why these observations were highlighted.

### Improved Version:

[http://www.math.usu.edu/~symanzik/teaching/2011\\_stat6560/RDataAndScripts/Yoshioka.R](http://www.math.usu.edu/~symanzik/teaching/2011_stat6560/RDataAndScripts/Yoshioka.R)

## 1.5 Rules for Good Data Displays

Wainer (1997), p. 46, suggests:

- “1. Examine the data carefully enough to know what they have to say, and then let them say it with a minimum of adornment.
2. In depicting scale, follow practices of “reasonable regularity.”
3. Label clearly and fully.”

Tufte (1983), p. 77, suggests:

“Graphical integrity is more likely to result if these six principles are followed:

- The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities represented.
- 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.
- Show data variation, not design variation.
- In time-series displays of money, deflated and standardized units of monetary measurements are nearly always better than nominal units.
- The number of information-carrying (variable) dimensions depicted should not exceed the number of dimensions in the data.
- Graphics must not quote data out of context.”

Robbins (2005), pp. 375–377, provides a “*Checklist of Possible Graph Defects*” in her Appendix A:

**“Can the reader clearly see the graphical elements?”**

- Do the data stand out? Are there superfluous elements?
- Are all graphical elements visually prominent?
- Are overlapping plotting symbols visually distinguishable?
- Can superposed data sets be readily visually assembled?
- Is the interior of the scale–line rectangle cluttered?
- Do data labels interfere with the quantitative data or clutter the graph?
- Is the data rectangle within the scale–line rectangle?
- Do tick marks interfere with the data?
- Do tick mark labels interfere with the data?
- Are axis labels legible?
- Are there too many tick marks?

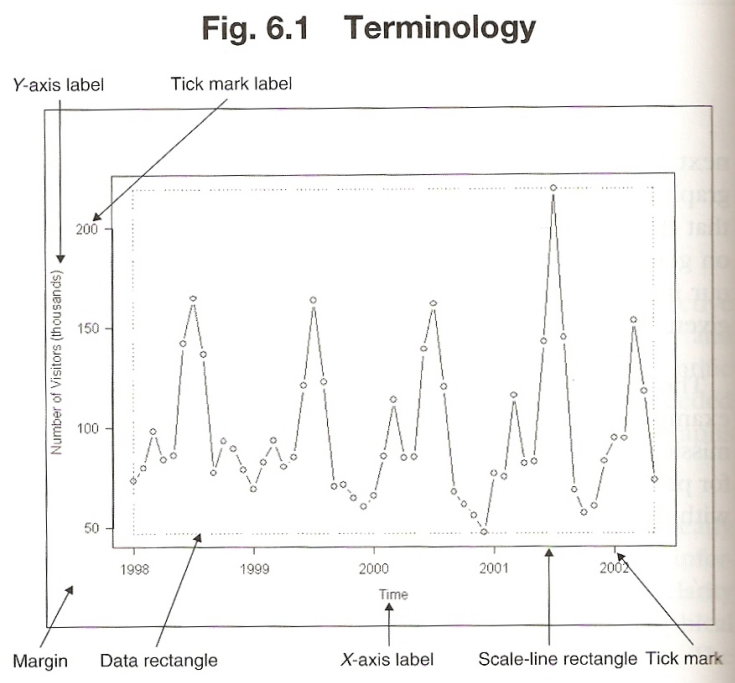


Figure 48: Robbins (2005), p. 156, Figure 6.1.

- Are there too many tick mark labels?
- Do the grid lines interfere with the data?
- Are there notes or keys inside the scale-line rectangle?
- Will visual clarity be preserved under reduction and reproduction?

**Can the reader clearly understand the graph?**

- Are the data drawn to scale?
- Is there an informative title?
- Is area or volume used to show changes in one dimension?
- Are there too many dimensions in the graph (more than in the data)?
- Are common baselines used wherever possible?
- Are all labels associated with the correct graphical elements?
- Is the reader required to make calculations?
- Are groups of charts drawn consistently?



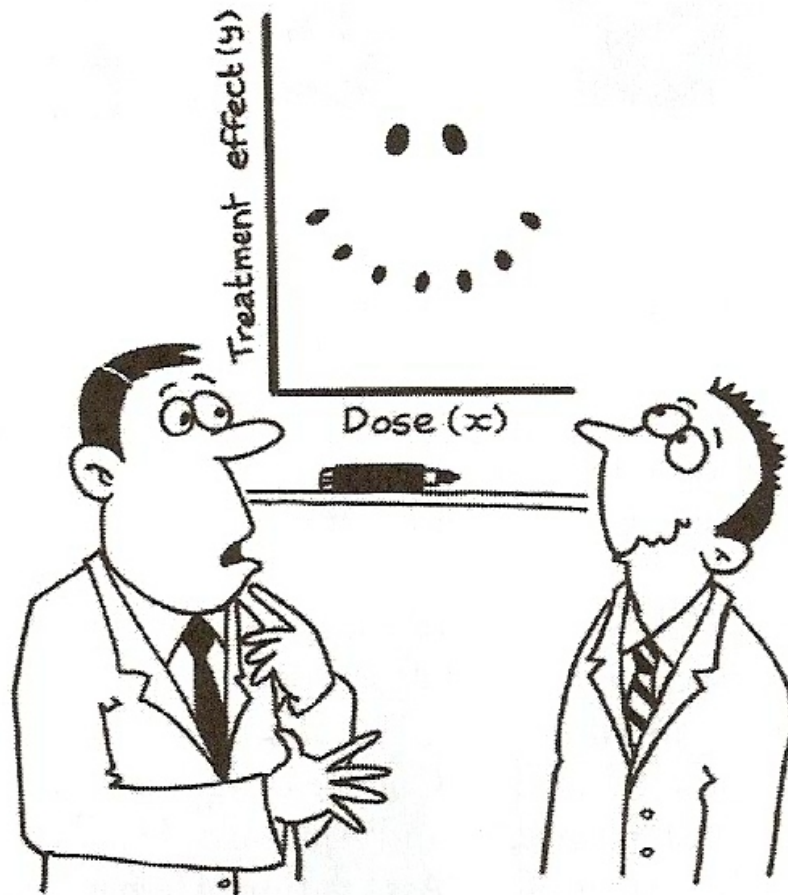
**Are the scales well chosen and labeled?**

- Is zero included for all bar graphs?
- Are there any unnecessary scale breaks?
- Is there a forceful indication of a scale break?
- Are there numerical values on two sides of a scale break that are connected?
- Does the aspect ratio allow the reader to see variations in the data?
- Are scales included for all axes?
- Are the scales labeled?
- Are tick marks at sensible values?
- Do the axes increase in the conventional direction?
- Does the data rectangle fill as much of the scale–line rectangle as possible?
- Are uneven time intervals handled correctly?
- Are the scales appropriate when different panels are compared?”

## 1.6 Further Reading

In addition to Wainer (1997), Tufte (1983), and Robbins (2005), cited so far in this chapter, many other sources exist that compare bad graphics with good graphics. Some of these additional sources are:

- Bertin (1977) and Bertin (2005) (first published in 1967)
- Henry (1995)
- Holmes (1991): check the author credentials and then decide whether this book is a source for good or bad graphics
- Huff & Geis (1954)
- Jones (2000)
- Kosslyn (1994) and Kosslyn (2006)
- Krämer (1991)
- Wainer (2005)
- Wainer (2007)
- Wallgren et al. (1996)
- Zelazny (2001)



**"It's a non-linear pattern with outliers.....but for some reason I'm very happy with the data."**

A CAUSE-commissioned cartoon that is part of the CAUSEweb collection and available for free noncommercial use by statistics teachers. Cartoon by John Landers ©. Provided by permission.

Figure 49: Amstat News, January 2009, p. 25, Cartoon.

## 2 History of Statistical Graphics: Plots, People, and Events

### 2.1 General History

- Michael Friendly's Web page: <http://www.math.yorku.ca/SCS/Gallery/milestone/>
- "*Milestones in the History of Data Visualization*" from the original "*Milestones in the History of Thematic Cartography, Statistical Graphics, and Data Visualization*". An illustrated chronology of innovations by Michael Friendly and Daniel J. Denis, York University, Canada. Organization by Mario Kanno (no longer active: [www.infografe.com.br](http://www.infografe.com.br); current: <http://kanno-infografia.blogspot.com/>). [http://www.math.yorku.ca/SCS/Gallery/milestone/Visualization\\_Milestones.pdf](http://www.math.yorku.ca/SCS/Gallery/milestone/Visualization_Milestones.pdf)
- "*Milestones in the History of Thematic Cartography, Statistical Graphics, and Data Visualization*". Michael Friendly, August 24, 2009. <http://www.math.yorku.ca/SCS/Gallery/milestone/milestone.pdf>

### 2.1.1 Milestones in the History of Data Visualization (According to Friendly)

**Pre-17th Century:** Early Maps and Diagrams

**1600–1699:** Measurement and Theory

**1700–1799:** New Graphic Forms

**1800–1849:** Beginnings of Modern Data Graphics

**1850–1899:** Golden Age of Data Graphics

**1900–1949:** Modern Dark Ages

**1950–1974:** Re-birth of Data Visualization

**1975–present:** High-D Data Visualization

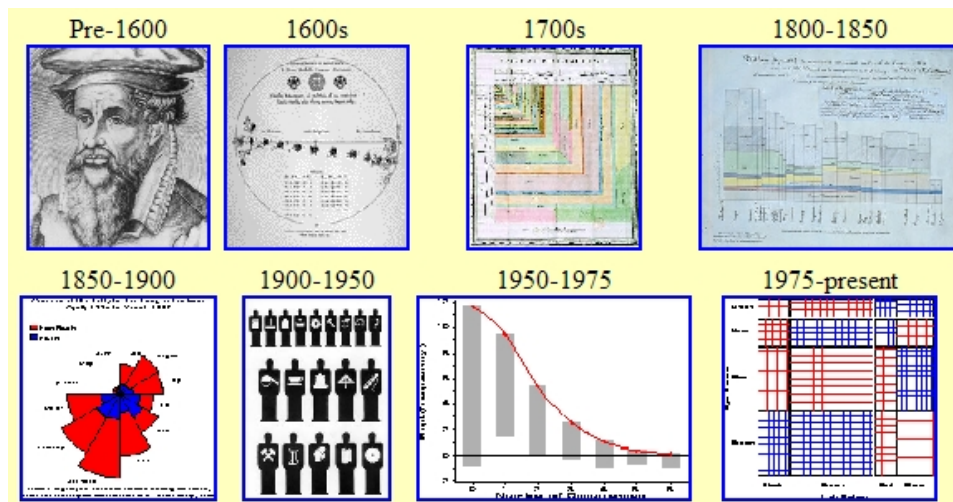


Figure 50: Screenshot taken from <http://www.math.yorku.ca/SCS/Gallery/milestone/> on 1/25/2009.

## 2.2 Selected People

Below are some of the individuals listed in Michael Friendly's *"Milestones in the History of Thematic Cartography, Statistical Graphics, and Data Visualization"*. Heyde & Seneta (2001) present biographies of 103 important statisticians born between 1601 to 1900. Christiaan Huygens, William Playfair, Florence Nightingale, and Francis Galton are listed in Heyde & Seneta (2001) as well as in Friendly's milestones overview.

**Christiaan Huygens:** (1629–1695), Netherlands

1669: First graph of a continuous distribution function, a graph of Gaunt's life table, and a demonstration of how to find the median remaining lifetime for a person of given age.

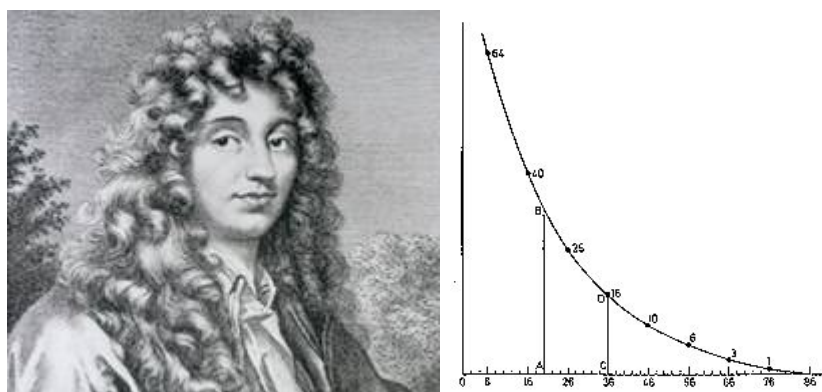


Figure 51: Figures taken from <http://www.math.yorku.ca/SCS/Gallery/images/portraits/huygens.jpg> and <http://www.math.yorku.ca/SCS/Gallery/images/huygens-graph.gif> on 1/27/2009.



**William Playfair:** (9/22/1759–2/11/1823), England

Symanzik et al. (2009), p. 552, state:

“Playfair is recognized as the *“man who invented outright the graphic method of representing statistical data”* (Funkhouser & Walker 1935, p. 103), the *“Apparent Inventor of Statistical Graphics”* (Funkhouser 1937, p. 280), the *“founder of graphic methods in statistics”* (FitzPatrick 1960, p. 39), the *“father of modern graphical display”* (Wainer 2005, p. 5), the *“progenitor of modern statistical graphics”* (Wainer 2005, p. 9), and the *“great pioneer of statistical graphics”* (Stephen Stigler, introducing Playfair 2005, opening page), to quote just a few of the accolades he received over the last 80 years.”

1786: Bar chart, line graphs of economic data.

1801: Invention of the pie chart, and circle graph, used to show part-whole relations.

Further Reading:

- Spence (2004)
- Spence (2006)
- Playfair (2005)
- [http://www.math.usu.edu/~symanzik/papers/2009\\_cost/editorial.html](http://www.math.usu.edu/~symanzik/papers/2009_cost/editorial.html)

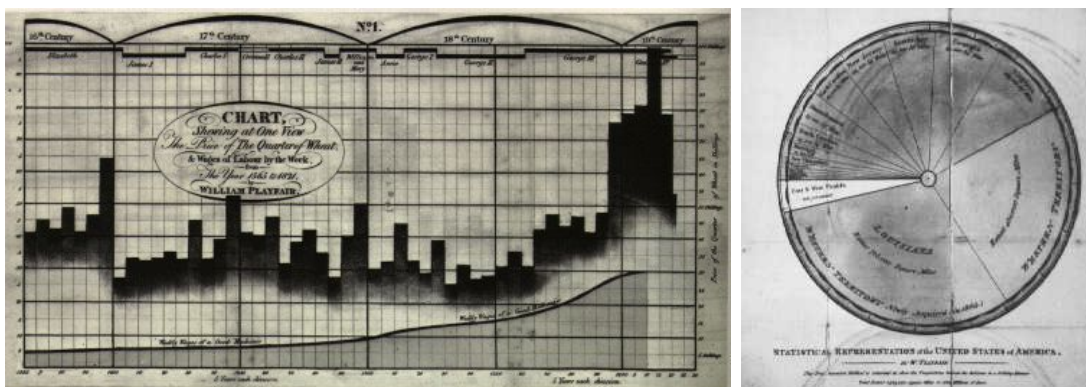


Figure 52: Figures taken from <http://www.math.yorku.ca/SCS/Gallery/images/playfair-wheat1.gif> and <http://www.math.yorku.ca/SCS/Gallery/images/playfair-pie.jpg> on 1/27/2009.

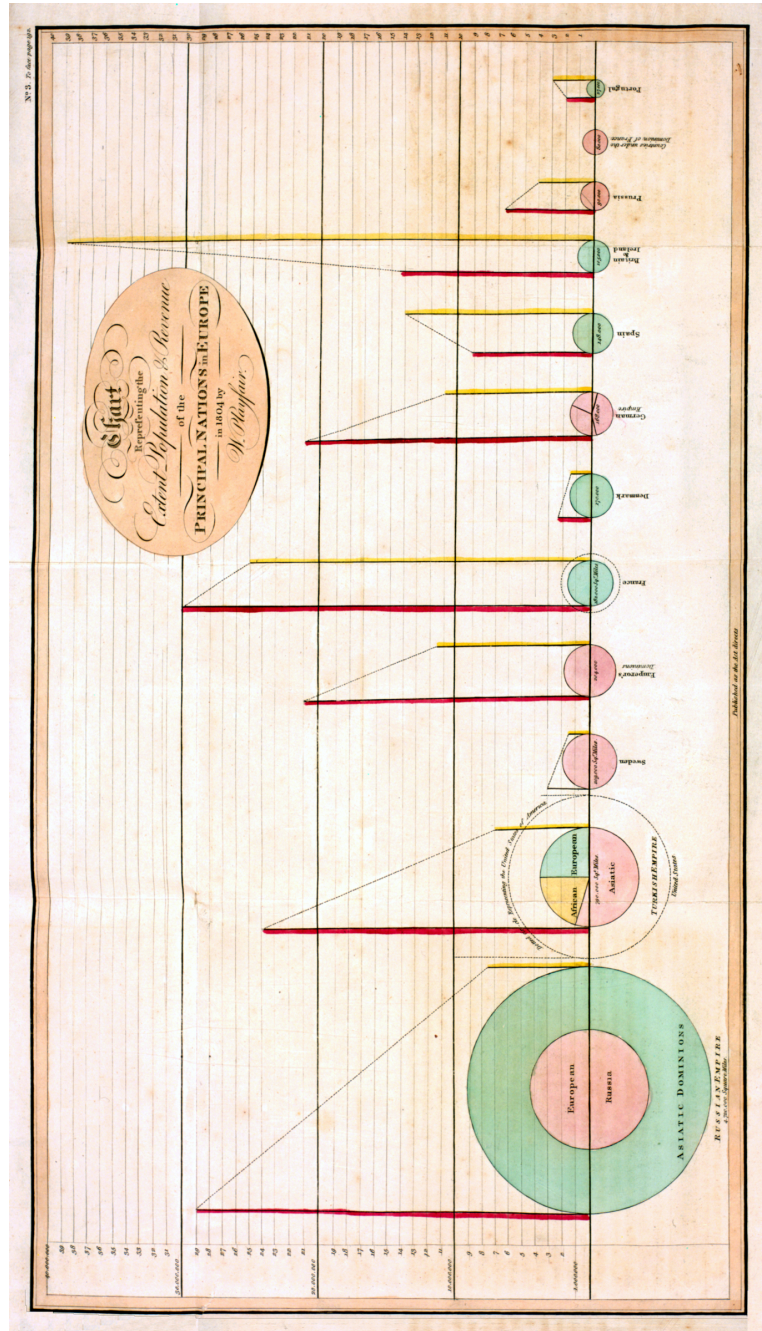


Figure 53: Symanzik et al. (2009), p. 553, Figure 1: *Chart Representing the Extent, Population & Revenue of the Principal Nations in Europe in 1804*. Plate 2 (labeled No. 3) from Playfair (1805). Courtesy of the Thomas Fisher Rare Book Library, University of Toronto.

**Charles Joseph Minard:** (1781–1870), France

1844: “Tableau-graphique” showing transportation of commercial traffic by variable-width (distance), divided bars (height  $\sim$  amount), area  $\sim$  cost of transport [An early form of the mosaic plot.]

1851: Map incorporating statistical diagrams: circles proportional to coal production (published in 1861).

1869: Minard’s flow map graphic of Napoleon’s Russian Campaign; often called “the best statistical graphic ever drawn” (Tufte 1983, p.40).

Further Reading:

- Tufte (1983), pp. 40–41 and more
- Wainer (1997), pp. 63–65
- Robinson (1967)
- Hankins (1999)
- Friendly (2000) (a pdf of the newsletter is available at <http://stat-computing.org/newsletter/v111.pdf>)

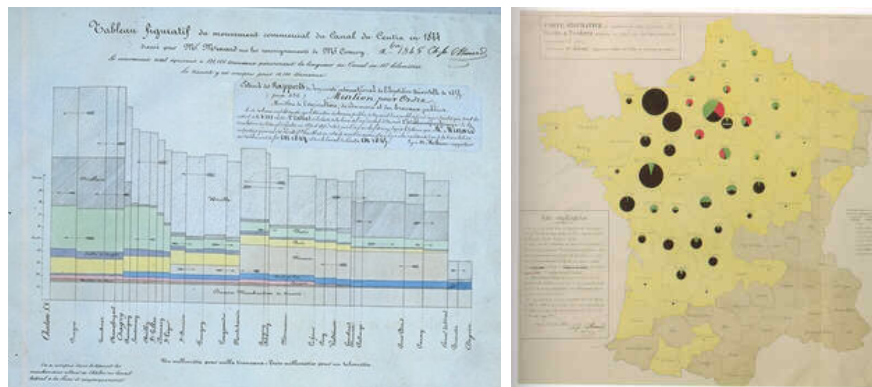


Figure 54: Figures taken from <http://www.math.yorku.ca/SCS/Gallery/images/enpc/img09a.jpg> and <http://www.math.yorku.ca/SCS/Gallery/images/Robinson/viandes.jpg> on 2/1/2009.

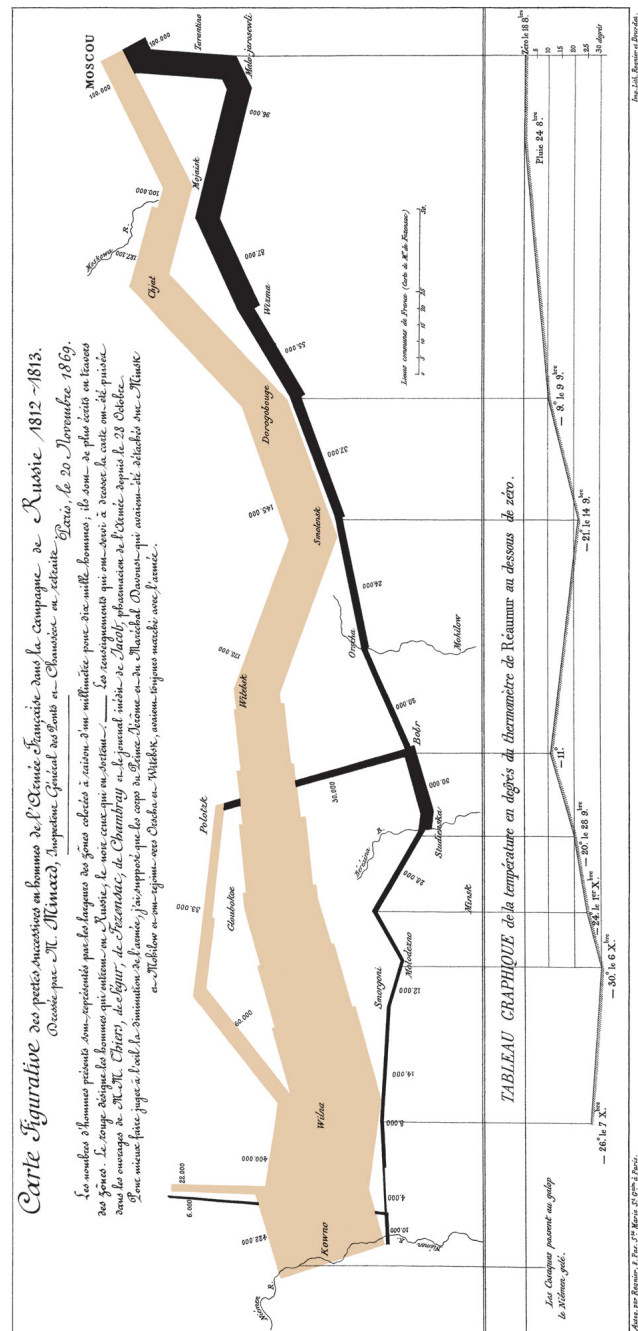


Figure 55: Figure taken from <http://cartographia.wordpress.com/2008/04/30/napoleons-invasion-of-russia/> on 2/8/2011.



**Florence Nightingale:** (1820–1910), England

1857: Polar area charts, known as “coxcombs” (used in a campaign to improve sanitary conditions of army) or as “Nightingale’s Rose”.

Additional details and an animation of her coxcombs can be found at [http://www.sciencenews.org/view/generic/id/38937/title/Math\\_Trek\\_\\_Florence\\_Nightingale\\_The\\_passionate\\_statistician](http://www.sciencenews.org/view/generic/id/38937/title/Math_Trek__Florence_Nightingale_The_passionate_statistician).

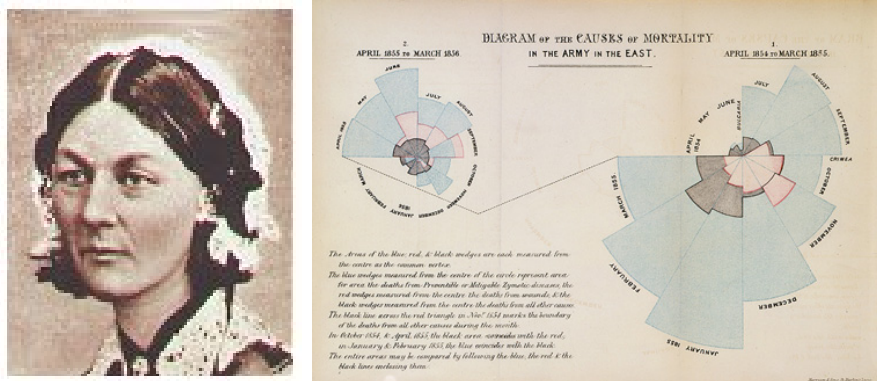


Figure 56: Figures taken from <http://www.math.yorku.ca/SCS/Gallery/images/portraits/nightingale.jpg> and <http://en.wikipedia.org/wiki/File:Nightingale-mortality.jpg> on 1/27/2009.



**Francis Galton:** (1822–1911), England

1861: The modern weather map, a chart showing area of similar air pressure and barometric changes by means of glyphs displayed on a map. These led to the discovery of the anti-cyclonic movement of wind around low-pressure areas.

c. 1874: Galton's first semi-graphic scatterplot and correlation diagram, of head size and height, from his notebook on *Special Peculiarities*.

1875: Galton's first illustration of the idea of correlation, using sizes of the seeds of mother and daughter plants.

1885: Normal correlation surface and regression, the idea that in a bivariate normal distribution, contours of equal frequency formed concentric ellipses, with the regression line connecting points of vertical tangents.

1899: Idea for “log-square” paper, ruled so that normal probability curve appears as a straight line.

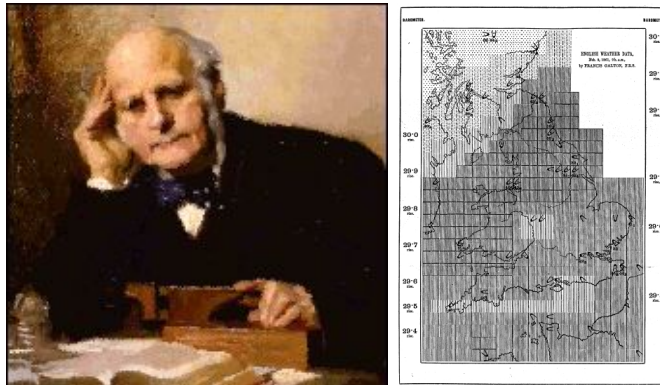


Figure 57: Figures taken from <http://www.math.yorku.ca/SCS/Gallery/images/portraits/galton-furse.gif> and <http://galton.org/essays/1860-1869/galton-1861-charts.pdf> on 1/27/2009.

**John W. Tukey:** (1915–2000), USA

1965: Beginnings of Exploratory Data Analysis (EDA): improvements on histogram in analysis of counts, tail values (hanging rootogram).

1969: Graphical innovations for exploratory data analysis (stem-and-leaf, graphical lists, box-and-whisker plots, two-way and extended-fit plots, hanging and suspended rootograms).

1974: Start of true interactive graphics in statistics; PRIM-9, the first system in statistics with 3-D data rotations provided dynamic tools for projecting, rotating, isolating and masking multidimensional data in up to nine dimensions — M. A. Fishkeller, Jerome H. Friedman and John W. Tukey.

1981: The “draftsman display” for three-variables (leading soon to the “scatter-plot matrix”) and initial ideas for conditional plots and sectioning (leading later to “coplots” and “trellis displays”) — John W. Tukey and Paul A. Tukey (a fifth cousin).

1990: Textured dot strips to display empirical distributions — Paul A. Tukey and John W. Tukey.

Further Reading:

- The PRIM-9 video, available at <http://stat-graphics.org/movies/>.
- Tukey (1977)
- Brillinger (2002) (preprint available at <http://www.stat.berkeley.edu/~brill/Papers/life.pdf>)

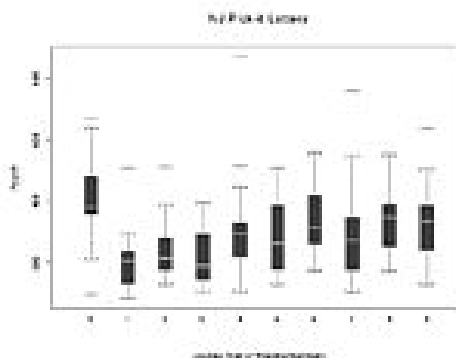


Figure 58: Figures taken from <http://www.math.yorku.ca/SCS/Gallery/images/portraits/tukey2.jpg> and <http://www.math.yorku.ca/SCS/Gallery/icons/NJPick-it.gif> on 2/1/2009.

**Jacques Bertin:** (1918–), France

1967: Comprehensive theory of graphical symbols and modes of graphics representation.

Among other things, Bertin introduced the idea of reordering qualitative variables in graphical displays to make relations more apparent, the reorderable matrix.

Further Reading:

- Bertin (1977)
- Bertin (2005)

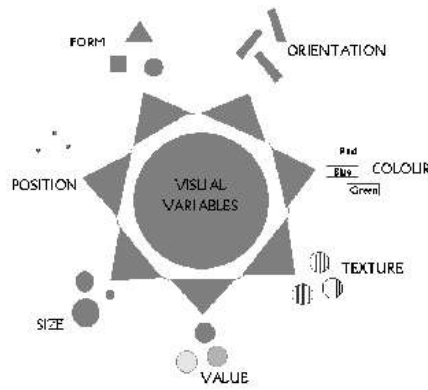


Figure 59: Figures taken from <http://www.math.yorku.ca/SCS/Gallery/images/portraits/jbertin.jpg> and <http://www.math.yorku.ca/SCS/Gallery/images/bertin-ve.jpg> on 2/1/2009.

## Andreas Buja and Collaborators: USA

Personal Home Page: <http://www-stat.wharton.upenn.edu/~buja/>

1988: First inclusion of grand tours in an interactive system that also has linked brushing, linked identification, visual inference from graphics, interactive scaling of plots, etc. — Andreas Buja, Daniel Asimov, Catherine Hurley and John A. McDonald.

1990: Grand tours combined with multivariate analysis — Catherine Hurley and Andreas Buja.

1991–1996: A series of developments and public distributions of highly interactive systems for data analysis and visualization (called XGobi) — Deborah Swayne, Di Cook and Andreas Buja.

2006: Buja's thoughts on Statistical Graphics: “[...] *theoretical foundations and math are important, right, but when I think back about what really may have had the most impact in what I did in the various labs that I worked, it's graphics! You know whenever I made a striking picture, people actually went “aahh,” “wow,” “that's great!”*”, “*Why don't we do more of this?*” *Pictures really, really speak.*” (Symanzik 2008, p. 182)

### Further Reading:

- Interview, published in Symanzik (2008).
- Numerous contributions to interactive and dynamic statistical software, summarized in Symanzik (2004).
- Several videos related to his work, collected at <http://stat-graphics.org/movies/>.

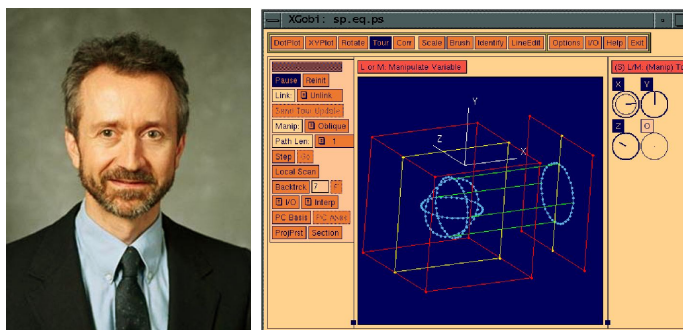
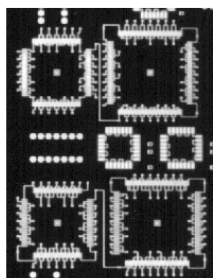


Figure 60: Figures taken from <http://www.math.yorku.ca/SCS/Gallery/people/AndreasBuja.jpg> and <http://www.research.att.com/areas/stat/xgobi/> on 2/1/2009.

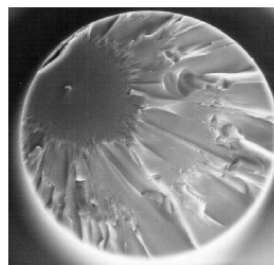
## John W. Chambers and Collaborators: USA

Personal Home Page: <http://stat.stanford.edu/~jmc4/>

1978: S, a language and environment for statistical computation and graphics. S (later sold as a commercial package, S-Plus; more recently, a public-domain implementation, R is widely available), would become a *lingua franca* for statistical computation and graphics — Richard A. Becker and John M. Chambers.



Wafer Solder Image



Optical Fiber Preform

Figure 61: Figures taken from <http://stat.stanford.edu/~jmc4/> and <http://stat.stanford.edu/~jmc4/papers/93.1.ps> on 2/1/2009.



**Antony Unwin and Collaborators:** Ireland, England, Germany

Personal Home Page: <http://stats.math.uni-augsburg.de/~unwin/>

1988: Interactive graphics for multiple time series with direct manipulation (zoom, rescale, overlaying, etc.) — Antony Unwin and Graham Wills.

1989: Statistical graphics interactively linked to map displays — Graham Wills, J. Haslett, Antony Unwin and P. Craig.

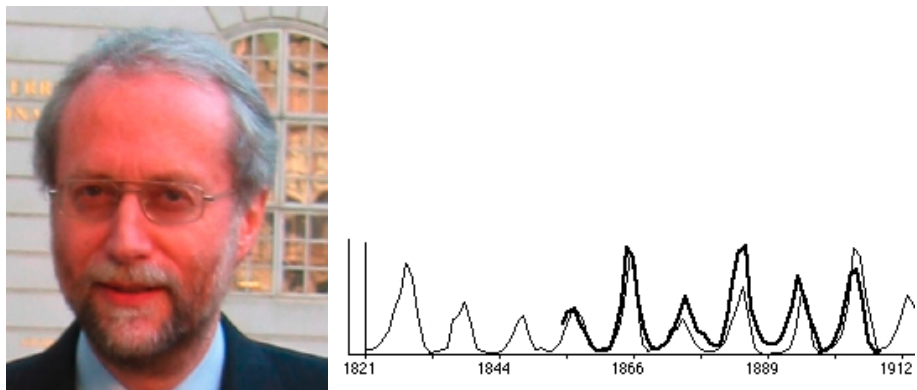


Figure 62: Figures taken from <http://stats.math.uni-augsburg.de/~unwin/> and <http://www.math.yorku.ca/SCS/Gallery/images/DiamondFast.jpg> on 2/1/2009.

**Edward J. Wegman: USA**

Personal Home Page: [http://statistics.gmu.edu/people\\_pages/wegman.html](http://statistics.gmu.edu/people_pages/wegman.html)

See here for a summary of his impressive vita — but also read about his denial of global warming:

<http://www.nationalpost.com/story.html?id=22003a0d-37cc-4399-8bcc-39cd20bed2f6&k=0>

1990: Statistical theory and methods for parallel coordinates plots.

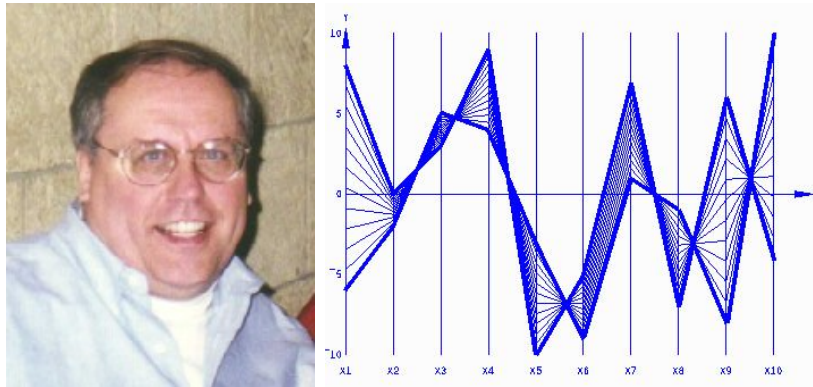


Figure 63: Figures taken from <http://www.math.yorku.ca/SCS/Gallery/people/EdWegman.jpg> and <http://www.math.yorku.ca/SCS/Gallery/images/parallel-coords.gif> on 2/1/2009.

## 2.3 Statistical Graphics and Events in History

### 2.3.1 John Snow and the Cholera Epidemic in London, 1854

(Based on a Student Project by William L. Welbourn in Spring 2009)

- John Snow (1813 – 1858): British anesthesiologist
- His investigation of the 1854 cholera outbreak in London, pioneered the field of epidemiology
- Some consider him the “*father of epidemiology*” (Vachon 2005)
- Cholera (Centers for Disease Control and Prevention 2010):
  - Acute diarrheal illness caused by intestinal infection by the bacterium *Vibrio cholerae*
  - Leads to rapid loss of body fluids and ultimately to dehydration and shock
  - Without treatment, death can occur within hours
- 1854 London Cholera Outbreak:
  - Six week period, beginning August 19, 1854
  - More than 575 deaths
  - “... Mortality in this limited area probably equals any that was ever caused in this country, even by the plague.” (Snow 1936)
- Snow’s Hypotheses:
  - Cholera is transmitted from person to person via fecal-oral route
  - Incubation period is 24 to 48 hours
  - The drinking water of the Broad Street Pump was the cause of the cholera outbreak



Figure 64: John Snow's rendition of the 1850 C.F. Cheffins Company Map, taken from [http://www.ph.ucla.edu/epi/snow/snowmap1\\_1854\\_lge.htm](http://www.ph.ucla.edu/epi/snow/snowmap1_1854_lge.htm). The white section of the map shows the area of particular interest to John Snow. The area encompassing the Broad Street Pump is circled (blue).

- Snow's Action to Control the Epidemic:
  - Utilized his map and empirical evidence to convince the Board of Guardians to remove the handle of the Broad Street Pump
  - A mere 48 fatal attacks occurred, following the removal of the handle of the Broad Street Pump, indicative that the water feeding the Broad Street Pump could be the source of the cholera epidemic

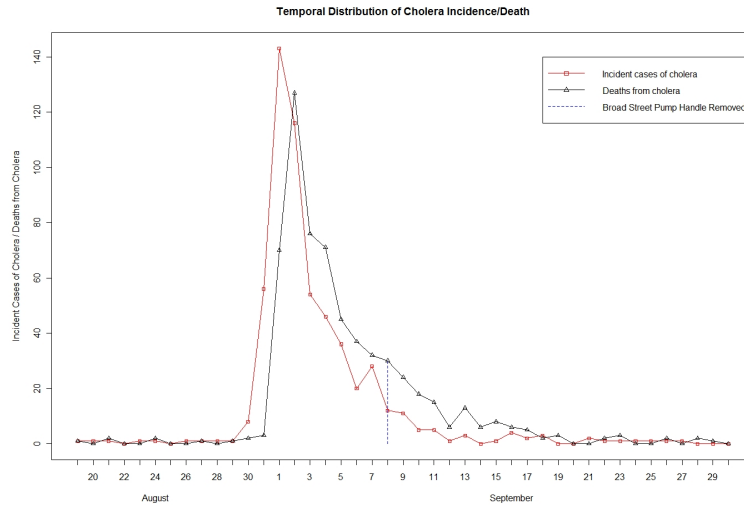


Figure 65: Time-series plot of the incident cases (red line) of cholera and deaths (black line) from cholera, for the time period, August 19, 1854 to September 30, 1854. The handle of the Broad Street Pump was removed September 8, 1854. (By William L. Welbourn)

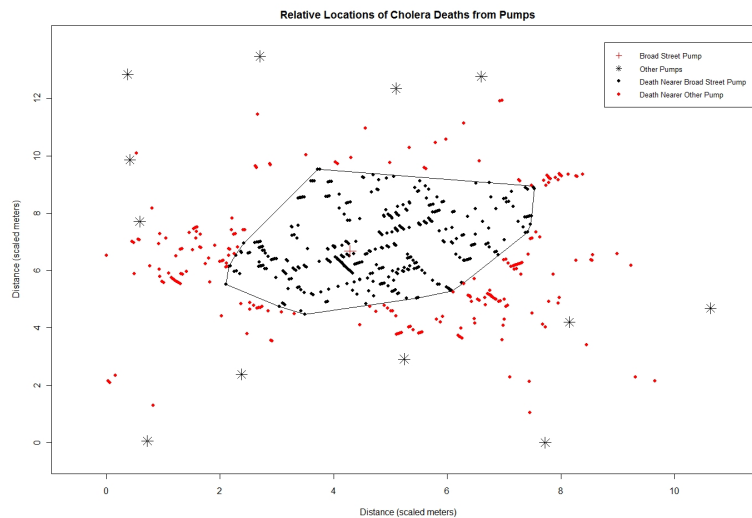


Figure 66: Relative positions of the 578 deaths arising from cholera and the thirteen pumps. 359 (62%) of these deaths occurred (black dots within the polygon) at a distance closer to the Broad Street Pump than to any other pump location. (By William L. Welbourn)

Further Reading:

- *John Snow — A Historical Giant in Epidemiology*, accessible at <http://www.ph.ucla.edu/epi/snow.html>
- Snow (1936), pp. 36–55
- Tufte (1997), pp. 27–37
- Wainer (1997), pp. 60–62
- Carvalho et al. (2004)
- R code to reproduce Figures 65 and 66:  
[http://www.math.usu.edu/~symanzik/teaching/2009\\_stat6560/RDataAndScripts/welbourn\\_william\\_project1\\_cholera.R](http://www.math.usu.edu/~symanzik/teaching/2009_stat6560/RDataAndScripts/welbourn_william_project1_cholera.R)



### 2.3.2 The Challenger Disaster, 1986

(Based on a Student Project by Abbass Sharif in Spring 2009)

- Background:
  - Lunch Time: January 28, 1986
  - The temperature was 31°F
  - Exploded after 73 seconds from its lunch leading to the death of its seven crew members
  - The Shuttle consisted of:
    - \* The orbiter: Housed crew and controls
    - \* An external fuel tank
    - \* Two solid-rocket booster motors
    - \* Each rocket-booster was shipped in 4 pieces
    - \* Each rocket-booster has three joints called *O-rings* (6 total)

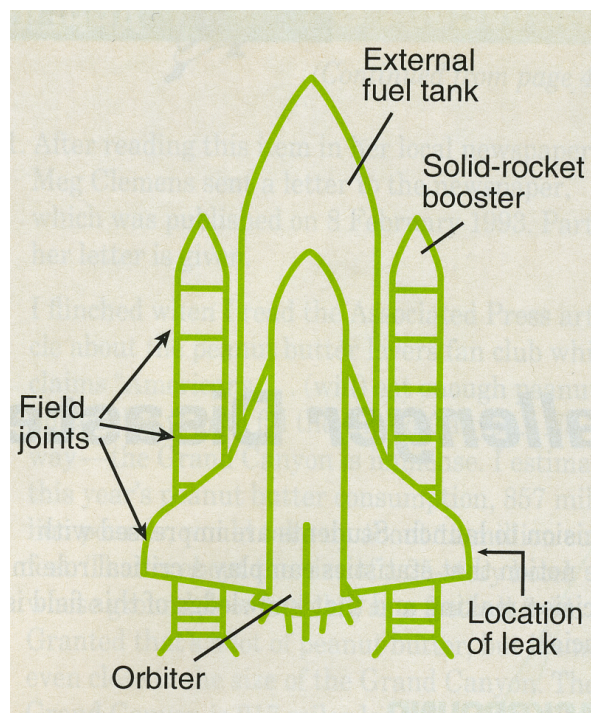


Figure 67: Figure taken from Tappin (1994).

- Challenger Pre-launch Discussion:

- The night before the scheduled launch, discussions occurred as to whether the launch should be postponed because of the low temperature
- It was believed, by some of the people who were involved in the decision, that low temperatures might harden the O-ring seals, and thus leading to a potentially dangerous combustion-gas leak (Wainer 1997)
- The O-rings had been designated as a “Criticality 1” component
- The engineers and manufacturers of the rocket motors believed that they should abort the flight
- They presented several (hand written) tabulated numbers to show their point

BLOW BY HISTORY		HISTORY OF O-RING TEMPERATURES (DEGREES - F)				
SRM-15 WORST BLOW-BY		MOTOR	MBT	AMB	O-RING	WIND
o 2 CASE JOINTS (80°), (110°) ARC		DM-1	68	36	47	10 MPH
o MUCH WORSE VISUALLY THAN SRM-22		DM-2	76	45	52	10 MPH
SRM 22 BLOW-BY		QM-3	72.5	40	48	10 MPH
o 2 CASE JOINTS (30-40°)		QM-4	76	48	51	10 MPH
		SRM-15	52	64	53	10 MPH
SRM-13A, 15, 16A, 18, 23A 24A		SRM-22	77	78	75	10 MPH
o NOZZLE BLOW-BY		SRM-25	55	26	29	10 MPH
					27	25 MPH

Figure 68: Tufte (1997), p. 42, Figure.

- These tables were unconvincing to their managers because:
  - \* NASA’s pressure
  - \* Did not show clearly the relationship between temperature and the number of O-rings failing to operate

- Challenger Post-launch Discussion:

- Subsequent to the explosion, commission staff members tried to graphically replicate the flaws in the pre-launch reasoning process
- They plotted the data from previous twenty three space shuttle launching where there was at least one O-ring failure
- The dataset consisted of two variables: the launching temperature and number of damaged O-rings

- The temperature has an average around 63°F and standard deviation equal to 8°F
- The conclusion was: there is no effect of temperature

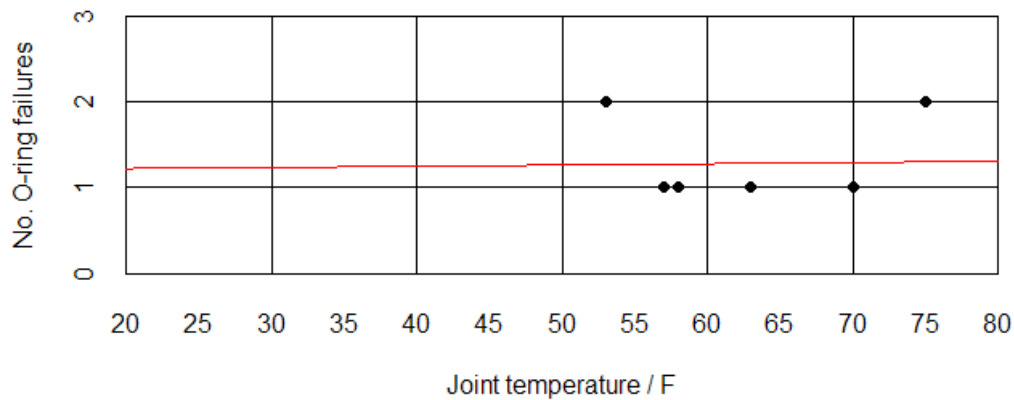


Figure 69: Reconstruction of Tufte (1997), p. 46, Figure (Right), with a fitted (almost) horizontal straight line. Such a figure was created after the accident to explain the poor reasoning in the pre-launch debate. (By Abbass Sharif)

- What went wrong?
  - **The data were graphed and analyzed out of context**
  - The dataset of O-ring damages was very small (7 cases only)
- What could have been done?
  - Use the complete dataset, and don't include only O-ring failure cases
  - Extend the “Number of Incidents” axis limit to 6
  - Fit a non-linear curve

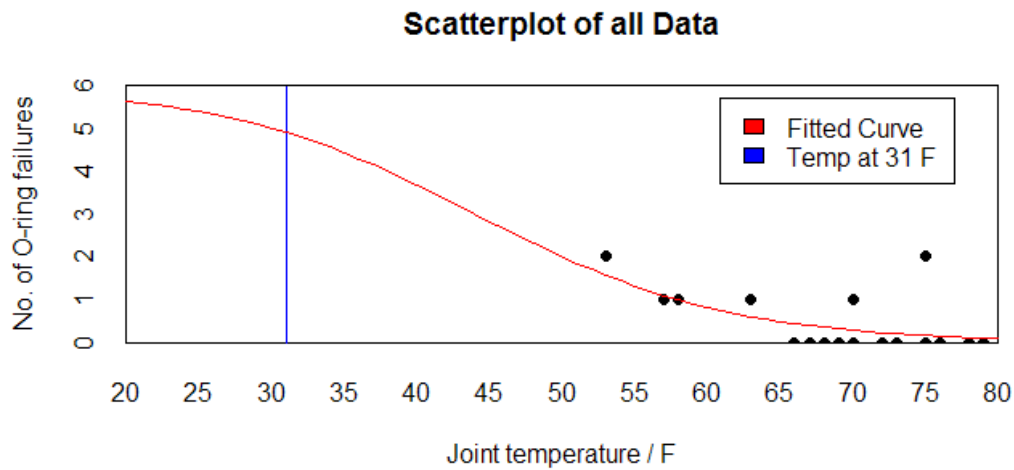


Figure 70: Reconstruction of a figure similar to Tufte (1997), p. 45, with a fitted non-linear curve. (By Abbass Sharif)

#### Further Reading:

- “Challenger Disaster Live on CNN”, Video posted at [http://youtube.com:  
http://www.youtube.com/watch?v=j4J0jcDFtBE](http://youtube.com:80/http://www.youtube.com/watch?v=j4J0jcDFtBE)
- Tappin (1994)
- Tufte (1997), pp. 27 & 38–53
- Wainer (1997), pp. 51–53
- R code to reproduce Figures 69 and 70:  
[http://www.math.usu.edu/~symanzik/teaching/2009\\_stat6560/RDataAndScripts/  
sharif\\_abbass\\_project1\\_challenger.R](http://www.math.usu.edu/~symanzik/teaching/2009_stat6560/RDataAndScripts/sharif_abbass_project1_challenger.R)

## 2.4 Further Reading

Additional sources for the history of statistical graphics, selected people, and events in history are:

- Friendly (2005) (3/15/2006 preprint available at <http://www.math.yorku.ca/SCS/Papers/gfkl.pdf>)
- Friendly (2008) (3/21/2006 preprint available at <http://www.math.yorku.ca/SCS/Papers/hbook.pdf>)
- Wainer (2009)

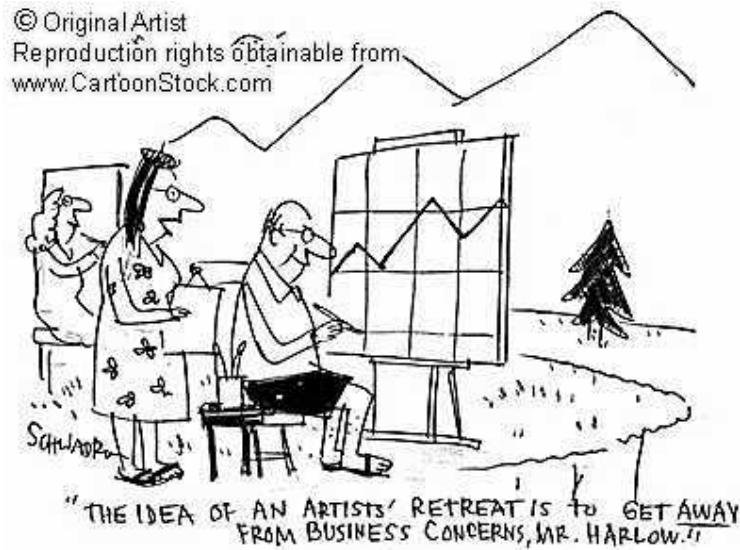


Figure 71: [http://www.cartoonstock.com/blowup\\_stock.asp?imageref=hsc3714&artist=Schwadron,+Harley&topic=statistics+](http://www.cartoonstock.com/blowup_stock.asp?imageref=hsc3714&artist=Schwadron,+Harley&topic=statistics+), Cartoon.

# Appendix



# Homework Assignments

## Homework Assignment 1 (1/12/2011)

15 Points — Due 1/21/2011 (pdf file via e-mail by 12noon)

(i) (15 Points — Individual Work) We will work with data from the Deepwater Horizon oil spill as one of our main data sets this semester. Think of at least ten other disasters since 1900 where technology was involved (i.e., no wars or diseases), such as the Titanic, the Challenger disaster, or the Chernobyl nuclear reactor disaster.

- Create a diagram in R that shows a timeline on the horizontal axis and marks your ten (or more) disasters. Use google Images search for *timeline* to get some idea about effective (and less effective) timeline plots.

R packages such as *diagram* (<http://cran.r-project.org/web/packages/diagram/>) or *igraph* (<http://cran.r-project.org/web/packages/igraph/index.html>) may be helpful to create your timeline. Or, you can simply draw your timeline “manually” by adding text, line segments, etc. to an empty plot area in R. All figures for my Stat 6710/20 lecture notes were created this way. For example, the figure for Theorem 1.2.1 (iv) on page 6 of my Stat 6710 lecture notes, accessible at [http://www.math.usu.edu/~symanzik/teaching/2010\\_stat6710/lect\\_main\\_full.pdf](http://www.math.usu.edu/~symanzik/teaching/2010_stat6710/lect_main_full.pdf), has been created as follows:

```
#R code to produce the graphic for Theorem 1.2.1 iv
#
pdf("lect_theorem1.2.1_iv.pdf", height = 8, width = 8)
plot(c(0,1), c(0,1), xlab = " ", ylab = " ", type = "n", axes = F)
lines(c(0, 1, 1, 0, 0), c(0, 0, 1, 1, 0))
deg = seq(0, 2 * pi, length = 1000)
x = sin(deg)
y = cos(deg)
lines(0.25*x + 0.4, 0.25*y + 0.6)
lines(0.20*x + 0.7, 0.20*y + 0.4)
text(0.4, 0.6, "A", cex = 3)
text(0.7, 0.4, "B", cex = 3)
title("Theorem 1.2.1 (iv)", cex.main = 3)
dev.off()
```

- Create two different plots that show the number of human deaths for your ten (or more) selected disasters. Sort your data in different ways and possibly transform the number of deaths (e.g., by taking the log). Provide meaningful titles and labels for your plots.
- Write a report in  $\text{\LaTeX}$  and translate via `pdflatex`. You should include an informative title page with your name, course information, and homework information.

Your main answer should contain a list of your ten (or more) disasters with clickable links (use `hyperref`) to an online reference where the disaster is described in more details.

Include your three figures (either in pdf or jpg format) into your  $\text{\LaTeX}$  document. Add a meaningful caption for each figure. Create some main text that briefly describes each of your figures and that links to these figures. Compare your two figures that show the number of human deaths and explain which of these is better (in your opinion).

Include your R code in the appendix (in a verbatim environment).

- Submit the single pdf file that is resulting from your work via e-mail to `symanzik@math.usu.edu` by 12noon on Fr 1/21/2011.
- Hint: If some of the expressions above mean nothing to you, google for them. Almost all documentation for R,  $\text{\LaTeX}$ , or any other software is nowadays freely accessible on the Web. It may take some time to search, but eventually you should find it! Also, if you have an idea how a graphic should look like, but have no idea how to get started, you may find it at the *R Graph Gallery* (<http://addictedtor.free.fr/graphiques/>). This Web site shows numerous graphics and it provides the R code how each graphic has been produced.

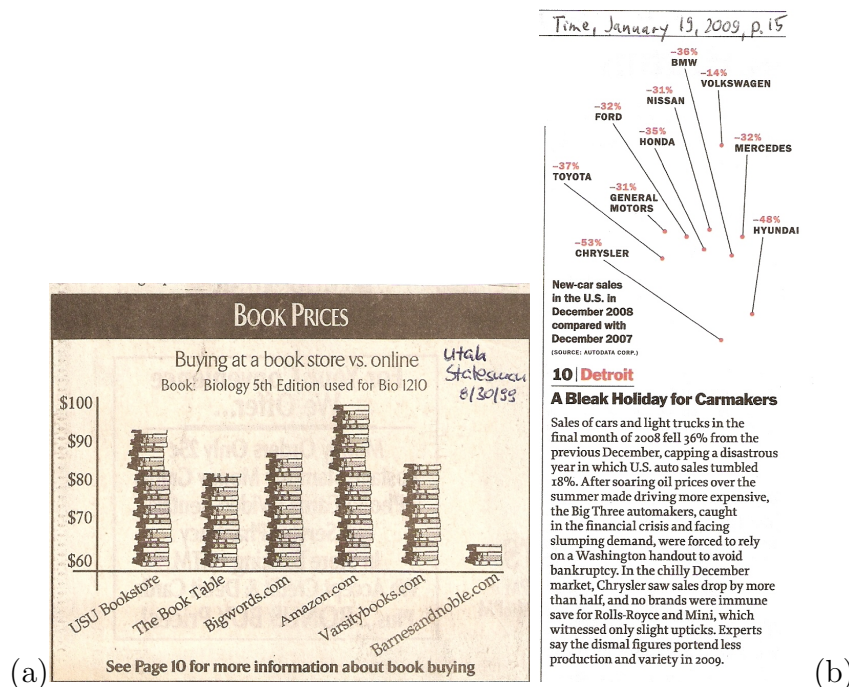
## Homework Assignment 2 (2/1/2011)

30 Points — Due 2/11/2011 (pdf file via e-mail by 5pm)

- (i) (8 Points — Group Work) Read Chapters 1, 2, and 3 of Tufte (1983) “*The Visual Display of Quantitative Information*”. Then take a closer look at the figures on top of p. 55 (“Comparative Annual Cost ...”) and on top of p. 69 (“Accroissement ...”).

(\*) For each of these figures, explain which rule(s) (how to construct a bad graphic) from our lecture notes the graphics designer has followed, i.e., list the rule(s) and explain why it has been followed. Demonstrate how these poor graphics might be improved. Using the data from the graphic (or your best approximation if necessary), construct a superior representation of the same information, using R. Include a short write-up (about half a page to a page) as to how you believe your version improves on the poor original.

- (ii) (8 Points — Group Work) Repeat (\*) for the two graphics included below.



- (iii) (8 Points — Group Work) Update your timeline plot from Homework 1 with events related to the Deepwater Horizon oil spill. Start with the actual explosion, the sinking, major attempts to stop the oil leakage, and the successful capping of the leak. The Data Expo 2011 Web page (<http://streaming.stat.iastate.edu/dataexpo/2011/>) may be a good starting point, but you should search for additional references beyond those listed on the Data Expo page. Also arrange (and summarize) the events from your timeline plot in a textual and/or tabular way.

When we take a closer look later at the actual data provided at the Data Expo 2011 Web page, we will see that the data are very heterogeneous. The events above may explain some of the sudden changes. There are likely other (environmental, climatic, etc.) variables that could lead to sudden changes. Hurricanes could be one such source. Find data related to hurricanes (since April 2010) that might have had an impact on the oil spill. These could be graphics (with dates) of hurricane paths, but even better, exact data with geographic locations of the hurricane centers. Document these data sources and download such data to your own computer. If there are other (environmental, climatic, etc.) variables that could have had an impact on the oil spill, do the same for these variables. Provide a brief textual summary of these additional data.

- (iv) (6 Points — Individual Work, due 2/18/2011, 5pm) Find a fresh example of a poor statistical graphic. Do not choose your example from one of the books for this class, or another book that specifically relates to graphics and charts, but from an original (preferably recent) source. Journal articles, newspapers, magazines, and scholarly books are all appropriate sources.

Each student must use a different graphic for this question. In fact, only the first person who notices a poor graphic can work on that particular graphic. If you notice a poor graphic, you have to send an e-mail to me, indicating something similar to the following: *"I found a poor graphic on page 1 of the Salt Lake Tribune on Wed 2/2/2011, titled ... that deals with ...."* Your bad graphic must meet at least three rules for bad graphics from Chapter 1. Good sources for bad graphics are CNN, Time magazine, the Utah Statesman, Wikipedia, and many other online sites, but also textbooks and journal papers.

Repeat (\*) for your graphic. Include the original (electronic) figure or turn in a scan or high quality photo of the original together with a precise reference (source and page, URL, etc.) where you found that figure.

## General Submission Rules: (for group and individual homework, reports, etc.)

All your submissions this semester must be typeset in  $\text{\LaTeX}$ . In fact, your submissions should translate via `pdflatex`. Figures (from scans or from graphical software) must accompany your  $\text{\LaTeX}$  document in electronic format. R code and data sets must be directly accessible from your document. You should assume that all documents reside in the same directory. For testing, one student should finalize all documents while another student checks the intended submission for completeness on a different computer.  $\text{\LaTeX}$  warnings are OK, but  $\text{\LaTeX}$  error messages will result in point deductions (depending on how much effort it takes on my side to fix a problem). Submit your files via e-mail to `symanzik@math.usu.edu`. In case your submission consists of four or more individual files, you have to collect these in a zip file and just submit this single zip file.

You will be allowed one original submission plus one revision where you should make changes, based on my feedback. For the first submission, just submit your resulting pdf file. For the second submission, do not submit a pdf file as I will retranslate all documents on my side. There will likely be opportunities for bonus points that will be awarded to the best group (or best individual) submission.

Your files should be named as follows (or in a closely related way):

```
groupI_hwJ_main.tex
groupI_hwJ_figK.pdf
groupI_hwJ_qL.R
groupI_hwJ_qM_data.xxx
```

```
lastname_firstname_hwJ_qL_main.tex
lastname_firstname_hwJ_qL_figK.pdf
```

```
lastname_firstname_projectN_main.tex
lastname_firstname_projectN_figK.pdf
```

where I, J, K, L, M, and N will be replaced by appropriate integers. xxx can be any acceptable extension for R data files. Include comments in your files where possible, e.g., dates, names, purpose of a file, etc. Also, include a `Readme.tex` that provides detailed instructions how I have to recreate a pdf file on my side.



Overall, please follow the general submission rules from Homework 1. In particular, please follow these instructions:

- Include a title page, including course name, student names, number of homework, and the submission date.
- Link your figures, tables, sections, R code, etc. within your document. Include meaningful figure and table captions. Provide clickable links to outside sources.
- Start answers to each question on a new page, clearly labeling which question is being answered. To combine multiple tex files created by different students, work with `\input`. Check that your overall document is consistent in appearance once you have combined individual tex files, e.g., consistent labeling, referencing, use of titles and headlines, etc.
- Include all R code in appendices so that I can run this code on my side. Test your code so that it runs correctly when copied from the pdf back into R. Incorrect R code will result in point deductions. I will sample some R code from the original submission, but will do some more careful testing only for the second submission.
- Include a References section. Make sure to include printed references, online references, and sources that inspired your R code. Do not forget to cite R and R packages you have used.

## References

- Anscombe, F. J. (1973), ‘Graphs in Statistical Analysis’, *The American Statistician* **27**(1), 17–21.
- Bertin, J. (1977), *La Graphique et le Traitement Graphique de l’Information*, Flammarion, Paris, France.
- Bertin, J. (2005), *Sémiologie Graphique: Les Diagrammes — Les Réseaux Les Cartes (4e édition)*, Les ré-impressions des Éditions de l’École des Hautes Études en Sciences Sociales, Paris, France.
- Brillinger, D. R. (2002), ‘John W. Tukey: His Life and Professional Contributions’, *The Annals of Statistics* **30**(6), 1535–1575.
- Carvalho, F. M., Lima, F. & Kriebel, D. (2004), ‘RE: On John Snow’s Unquestioned Long Division’, *American Journal of Epidemiology* **159**(4), 422.
- Centers for Disease Control and Prevention (2010), Cholera. Retrieved February 13, 2011 from <http://www.cdc.gov/cholera/index.html>.
- FitzPatrick, P. J. (1960), ‘Leading British Statisticians of the Nineteenth Century’, *Journal of the American Statistical Association* **55**(289), 38–70.
- Friendly, M. (2000), ‘Re-Visions of Minard’, *Statistical Computing and Statistical Graphics Newsletter* **11**(1), 1 & 13–19.
- Friendly, M. (2005), Milestones in the History of Data Visualization: A Case Study in Statistical Historiography, in C. Weihs & W. Gaul, eds, ‘Classification: The Ubiquitous Challenge’, Springer, New York, NY, pp. 34–52.
- Friendly, M. (2008), A Brief History of Data Visualization, in C. Chen, W. Härdle & A. Unwin, eds, ‘Handbook of Data Visualization’, Springer, Berlin, Heidelberg, pp. 15–56 & 2 Color Plates.
- Funkhouser, H. G. (1937), ‘Historical Development of the Graphical Representation of Statistical Data’, *Osiris* **3**, 269–404.
- Funkhouser, H. G. & Walker, H. M. (1935), ‘Playfair and his Charts’, *Economic History* **3**, 103–109.

- Hankins, T. L. (1999), ‘Blood, Dirt, and Nomograms: A Particular History of Graphs’, *Isis* **90**(1), 50–80.
- Henry, G. T. (1995), *Graphing Data: Techniques for Display and Analysis*, Sage Publications, Thousand Oaks, CA.
- Heyde, C. C. & Seneta, E., eds (2001), *Statisticians of the Centuries*, Springer, New York, NY.
- Holmes, N. (1991), *Designer’s Guide to Creating Charts & Diagrams (Paperback Edition)*, Watson–Guptill Publications, New York, NY.
- Huff, D. & Geis, I. (1954), *How to Lie with Statistics*, W. W. Norton & Company, New York, NY.
- Jones, G. E. (2000), *How to Lie with Charts*, toExcel Press, Lincoln, NE.
- Kosslyn, S. M. (1994), *Elements of Graph Design*, W. H. Freeman and Company, New York, NY.
- Kosslyn, S. M. (2006), *Graph Design for the Eye and Mind*, Oxford University Press, New York, NY.
- Krämer, W. (1991), *So lügt man mit Statistik (3. Auflage)*, Campus Verlag, Frankfurt/Main, Germany.
- Playfair, W. (1805), *An Inquiry into the Permanent Causes of the Decline and Fall of Powerful and Wealthy Nations*, Greenland & Norris, London, U.K.
- Playfair, W. (2005), *The Commercial and Political Atlas and Statistical Breviary, Edited and Introduced by Howard Wainer and Ian Spence*, Cambridge University Press, New York, NY.
- Robbins, N. B. (2005), *Creating More Effective Graphs*, Wiley, Hoboken, NJ.
- Robinson, A. H. (1967), ‘The Thematic Maps of Charles Joseph Minard’, *Imago Mundi* **21**, 95–108.
- Rosling, H. & Johansson, C. (2009), ‘Gapminder: Liberating the x-axis from the Burden of Time’, *Statistical Computing and Statistical Graphics Newsletter* **20**(1), 4–7.

- Snow, J. (1936), *Snow on Cholera: Being a Reprint of Two Papers by John Snow, M.D. Together with a Biographical Memoir by B. W. Richardson, M.D. and an Introduction by Wade Hampton Frost, M.D.*, The Commonwealth Fund & Oxford University Press, New York, NY & London, U.K.
- Spence, I. (2004), Playfair, William (1759–1823), in H. C. G. Matthew & B. Harrison, eds, ‘Oxford Dictionary of National Biography’, Oxford University Press, Oxford, U.K. <http://www.oxforddnb.com/view/article/22370> (accessed 13 Aug 2009).
- Spence, I. (2006), William Playfair and the Psychology of Graphs, in ‘2006 JSM Proceedings’, American Statistical Association, Alexandria, VA, pp. 2426–2436. (CD).
- Symanzik, J. (2004), Interactive and Dynamic Graphics, in J. E. Gentle, W. Härdle & Y. Mori, eds, ‘Handbook of Computational Statistics — Concepts and Methods’, Springer, Berlin, Heidelberg, pp. 293–336.
- Symanzik, J. (2008), ‘Interview with Andreas Buja’, *Computational Statistics* **23**(2), 177–184.
- Symanzik, J., Fischetti, W. & Spence, I. (2009), ‘Editorial: Commemorating William Playfair’s 250th Birthday’, *Computational Statistics* **24**(4), 551–566.
- Tappin, L. (1994), ‘Applications: Analyzing Data Relating to the Challenger Disaster’, *Mathematics Teacher* **87**(6), 423–426.
- Tufte, E. R. (1983), *The Visual Display of Quantitative Information*, Graphics Press, Cheshire, CT.
- Tufte, E. R. (1997), *Visual Explanations: Images and Quantities, Evidence and Narrative*, Graphics Press, Cheshire, CT.
- Tukey, J. W. (1977), *Exploratory Data Analysis*, Addison Wesley, Reading, MA.
- Vachon, D. (2005), ‘Doctor John Snow Blames Water Pollution for Cholera Epidemic’, *Old News* **16**(8), 8–10.
- Wainer, H. (1997), *Visual Revelations: Graphical Tales of Fate and Deception from Napoleon Bonaparte to Ross Perot*, Copernicus/Springer, New York, NY.
- Wainer, H. (2005), *Graphic Discovery: A Trout in the Milk and Other Visual Adventures*, Princeton University Press, Princeton, NJ.

- Wainer, H. (2007), ‘Improving Data Displays: Ours and the Media’s’, *Chance* **20**(3), 8–15.
- Wainer, H. (2009), ‘A Centenary Celebration for Will Burtin: A Pioneer of Scientific Visualization’, *Chance* **22**(1), 51–55.
- Wallgren, A., Wallgren, B., Persson, R., Jorner, U. & Haaland, J.-A. (1996), *Graphing Statistics & Data: Creating Better Charts*, Sage Publications, Thousand Oaks, CA.
- Yoshioka, K. (2002), ‘KyPlot — A User-Oriented Tool for Statistical Data Analysis and Visualization’, *Computational Statistics* **17**(3), 425–437.
- Zelazny, G. (2001), *Say it with Charts: The Executive’s Guide to Visual Communication (Fourth Edition)*, McGraw-Hill, New York, NY.

## — THE END —

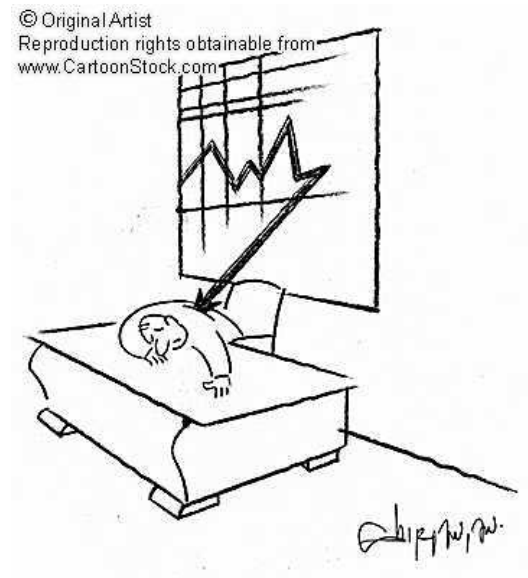


Figure 72: [http://www.cartoonstock.com/blowup\\_stock.asp?imageref=vsh0184&artist=Shirvanian,+Vahan&topic=statistics+](http://www.cartoonstock.com/blowup_stock.asp?imageref=vsh0184&artist=Shirvanian,+Vahan&topic=statistics+), Cartoon.