STAT 6560
Graphical Methods
Spring Semester 2009

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

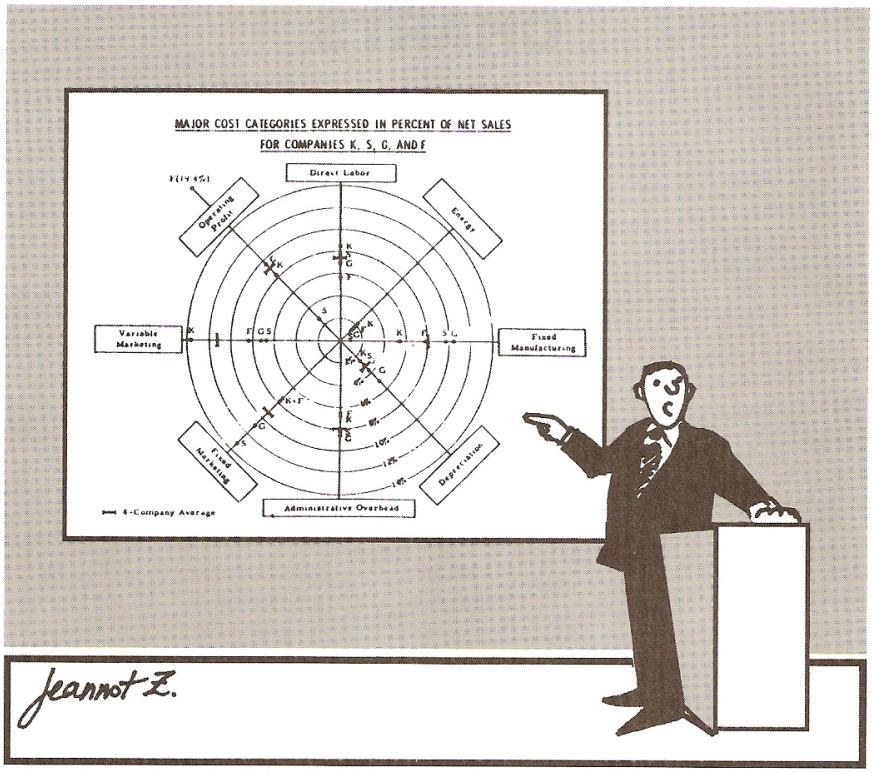
Web: http://www.math.usu.edu/~symanzik/

Contents

Acknowledgements

This course closely follows 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 will be 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://www.galaxy.gmu.edu/~dcarr/`). We are likely to include parts from additional Web sources that will be specified later on.

Jürgen Symanzik, January 6, 2009.



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

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

Lecture 02:
We 01/07/09

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 priciples will lead to a successful analysis and presentation.

1.2 Why Graphics ?!?

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

Start R and load the anscombe 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/2009_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/2009_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."[†]

Let us follow these strategies:

[†]Show the data unclearly.

1.3.1 Don't show much data

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

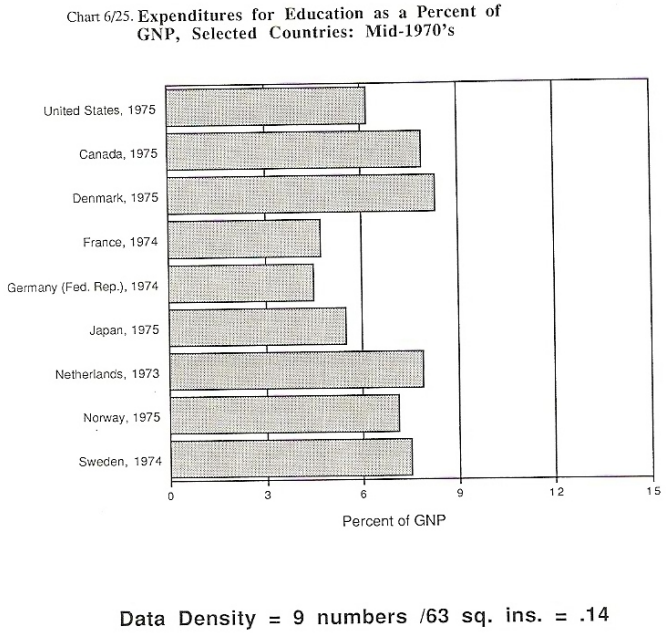


FIGURE 2. Chart 6/25 from *Social Indicators III* showing expenditures for education for nine countries as a function of GNP.

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

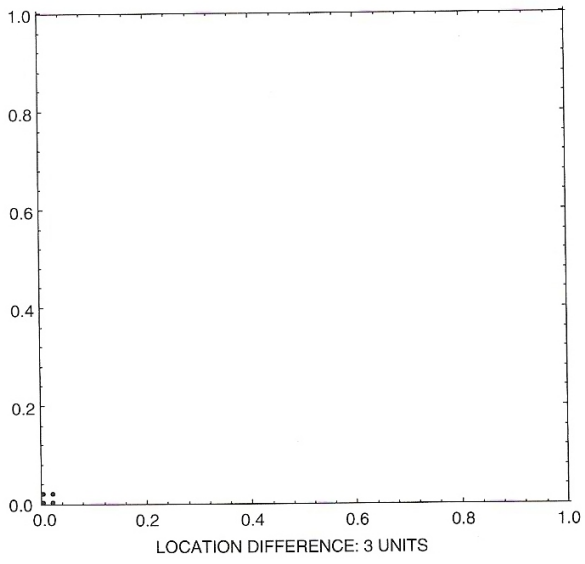


FIGURE 3. A graph of obviously low data density.

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

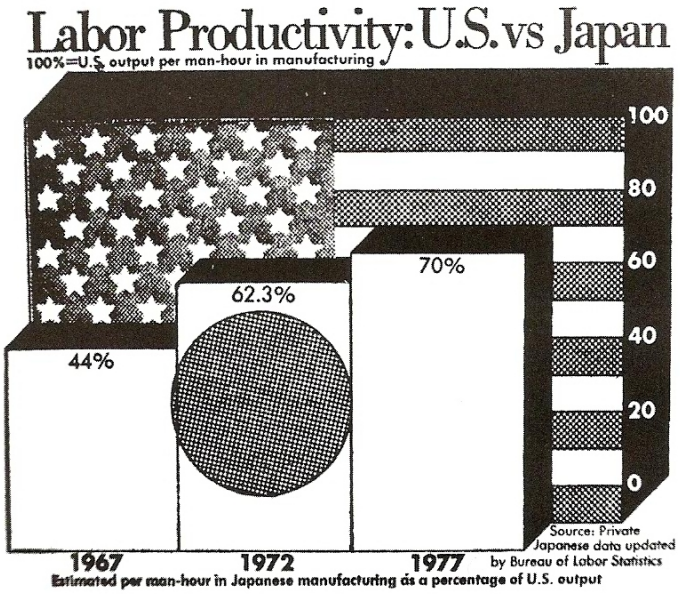


FIGURE 6. A graph with low data density filled in with chartjunk from the *Washington Post*, 1978.

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

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

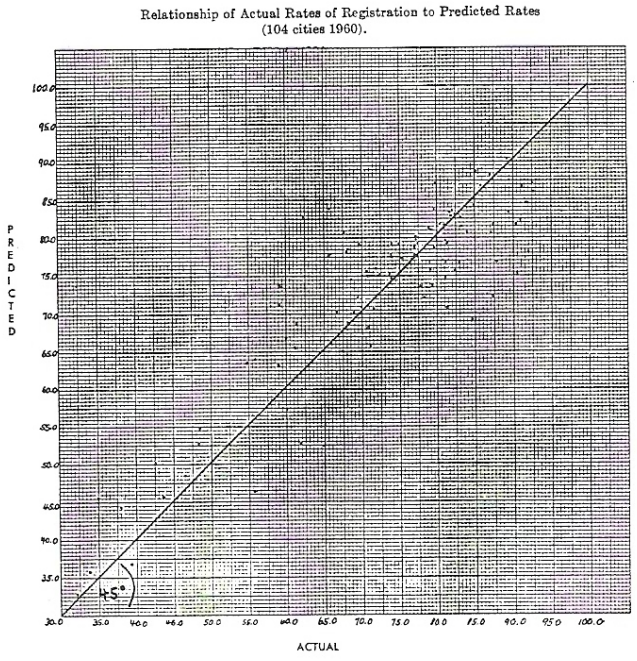


FIGURE 8. Hiding the data in the grid.

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

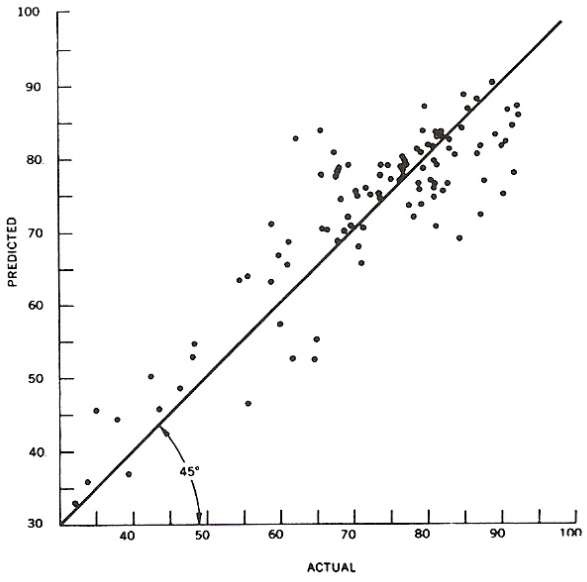
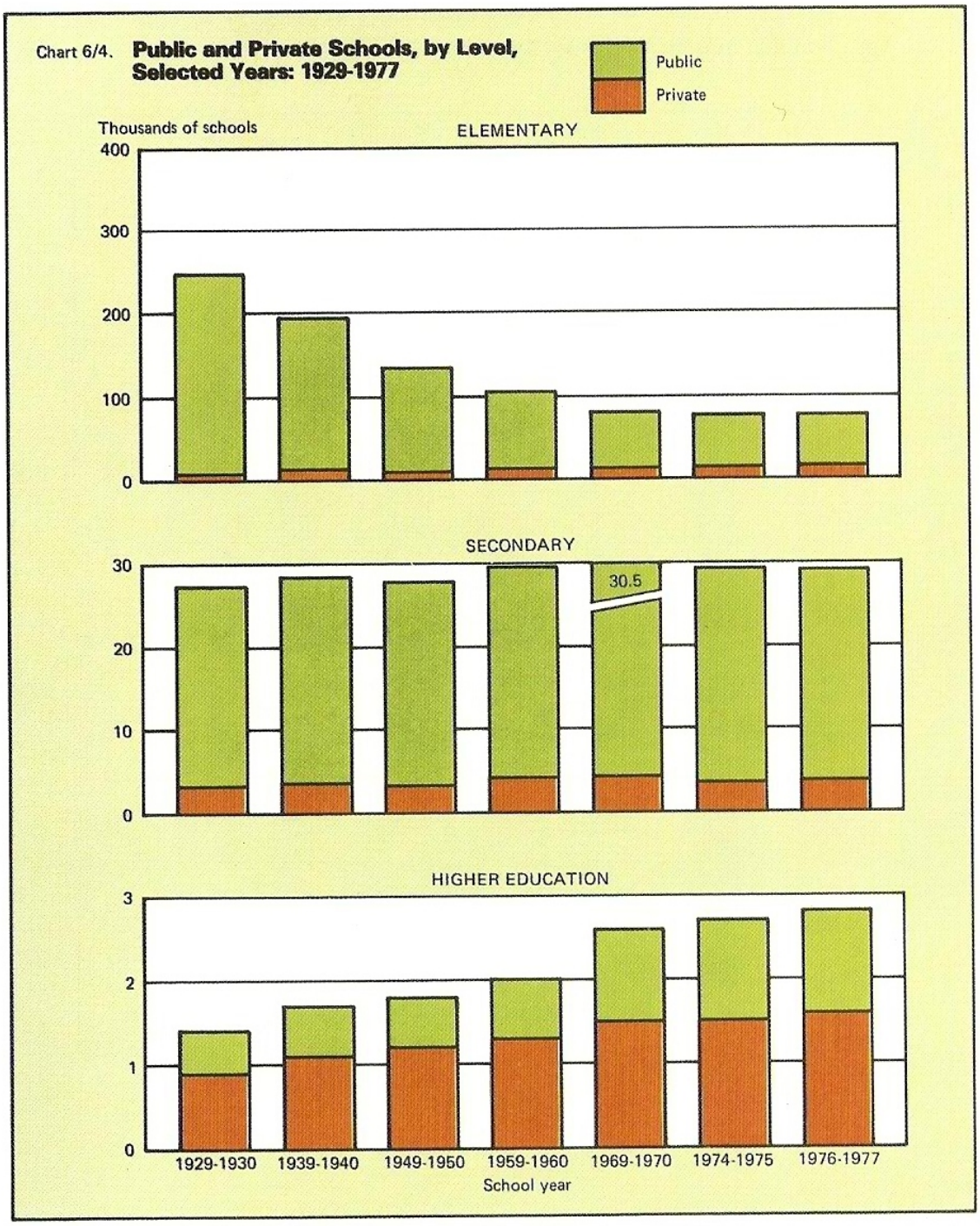


FIGURE 10. A redone example of the data from figure 8.

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.

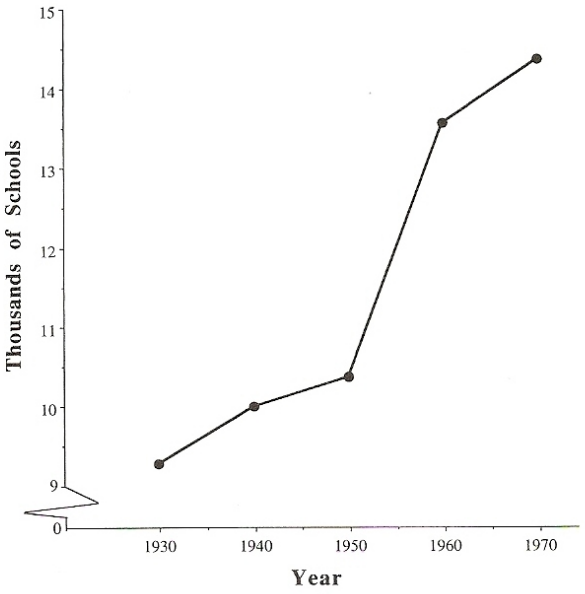


FIGURE 12. Expanding the scale and showing the data for the number of private elementary schools from figure 11.

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

1.3.2 Show the data inaccurately

Lecture 03:
Fr 01/09/09

Rule 3: Ignore the visual metaphor altogether.

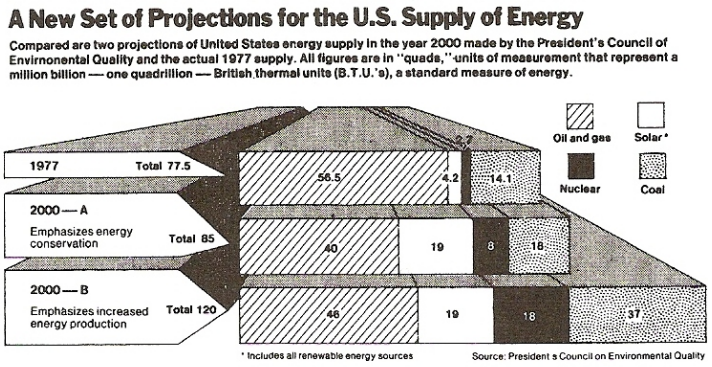


FIGURE 13. Ignoring the visual metaphor by letting a longer bar segment represent a smaller amount of coal (from the *New York Times*, 1978).

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

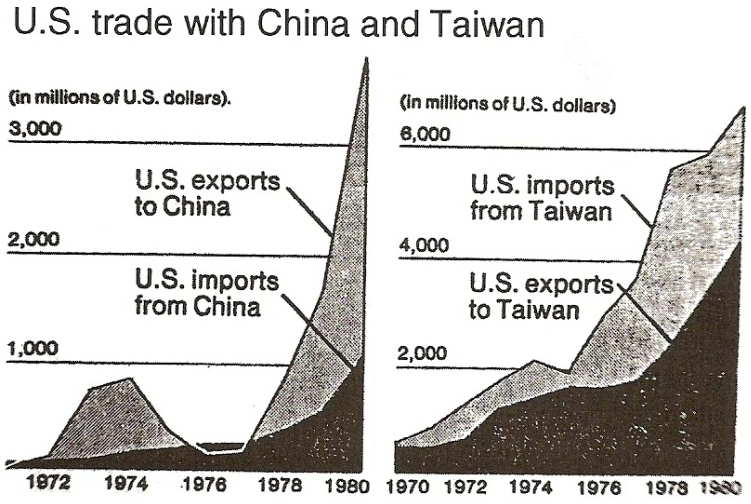


FIGURE 14. Reversing the metaphor in mid-graph while changing scales on both axes (from the *New York Times*, June 14, 1981).

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

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

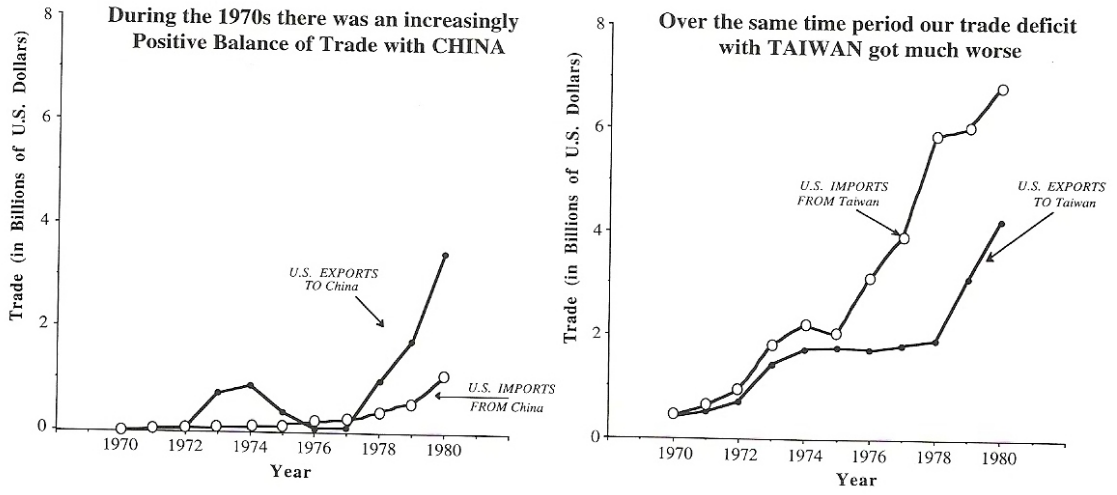


Figure 11: Wainer (1997), p. 21, Figure 15: Wainer (1997), p. 21, Figure 14, improved.

Rule 4: Only order matters.

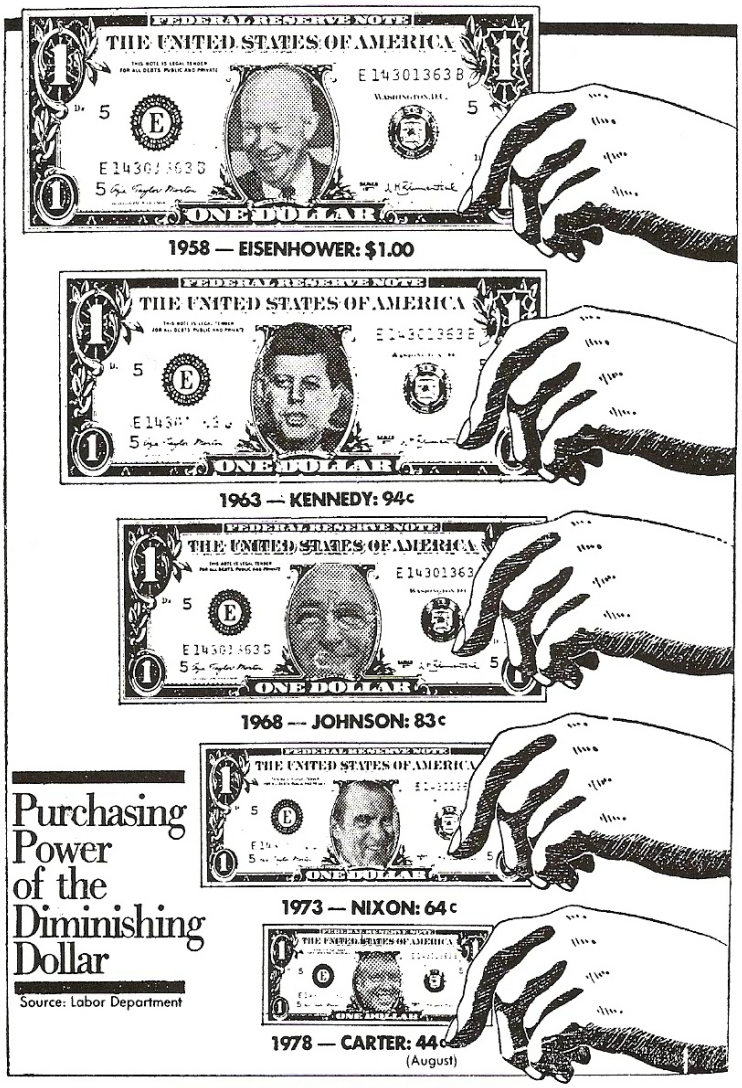


FIGURE 17. An example of how to goose up the effect by squaring the eyeball.

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

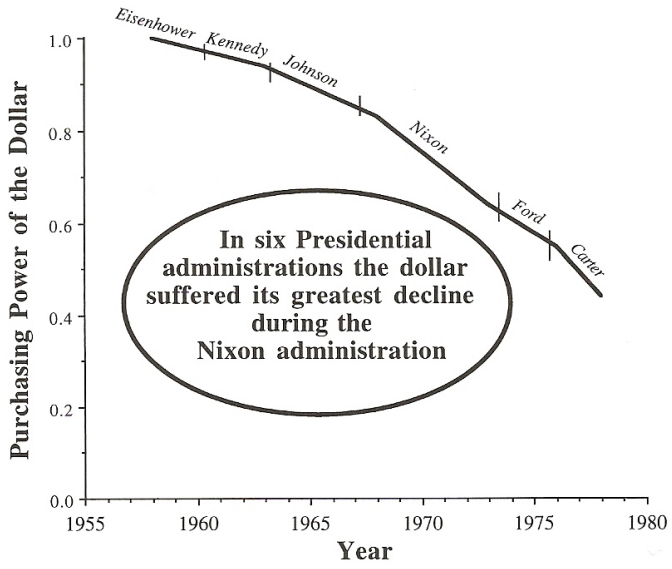


FIGURE 18. The data in figure 17 as an unadorned line chart (from Wainer, 1980).

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*).

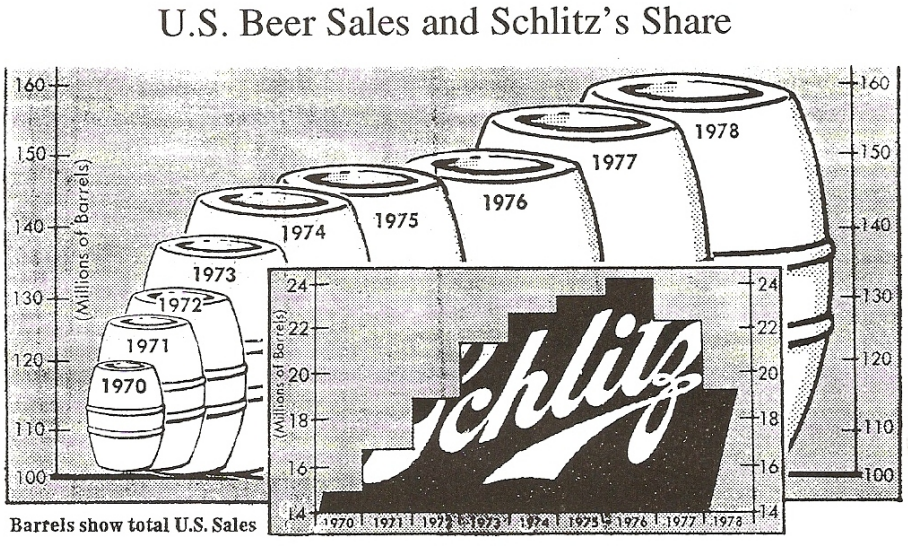


Figure 14: Wainer (1997), p. 24, Figure 19.

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.

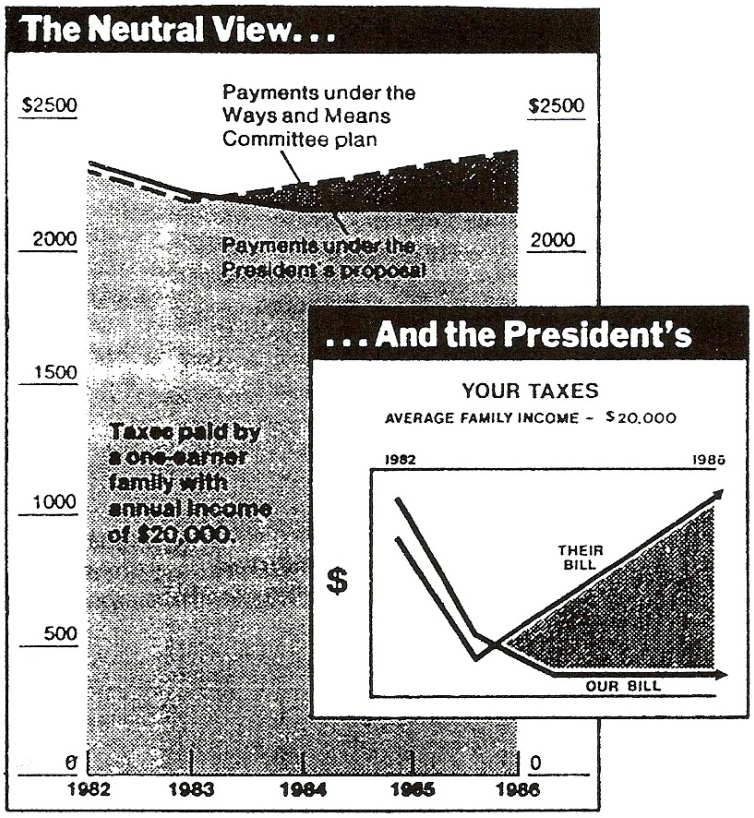


FIGURE 23. *New York Times* graphs showing how lack of context changes our perceptions about alternative tax bills.

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).

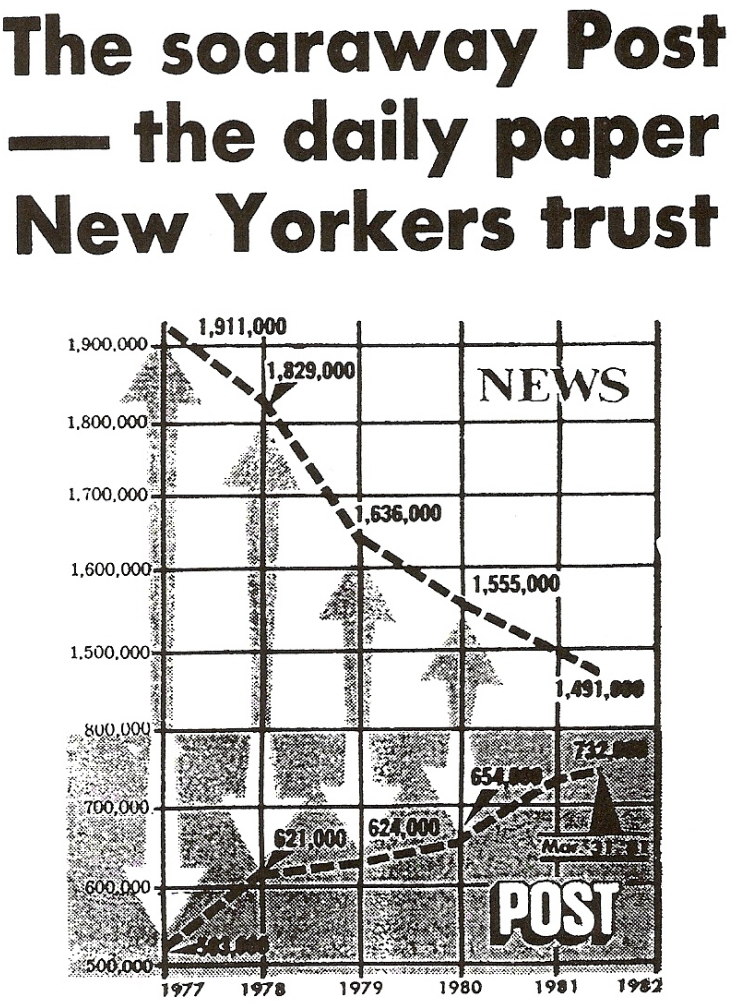


Figure 19: Wainer (1997), p. 28, Figure 24.

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).

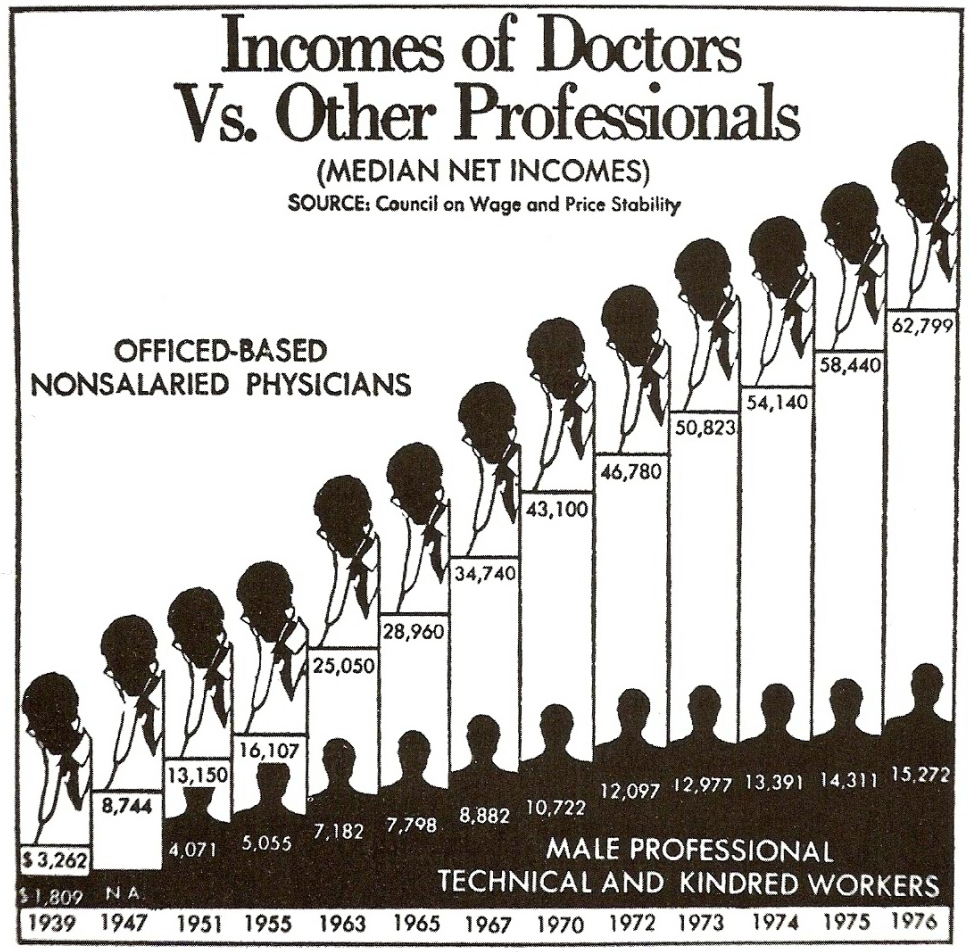


Figure 20: Wainer (1997), p. 29, Figure 25.

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

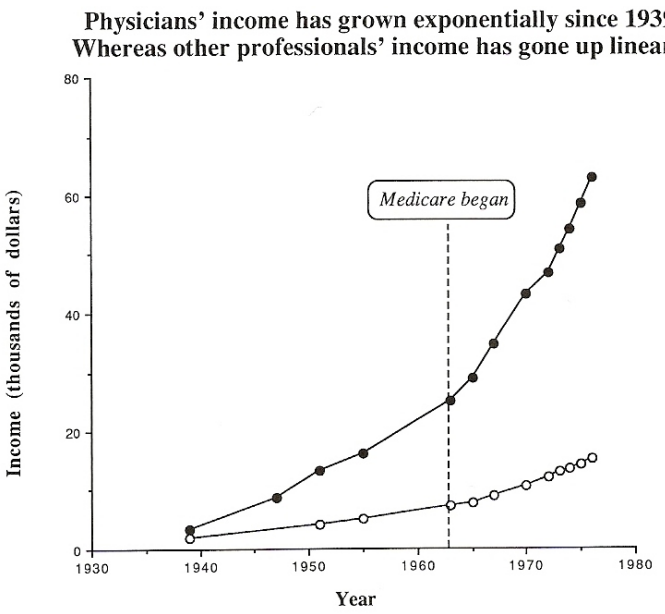
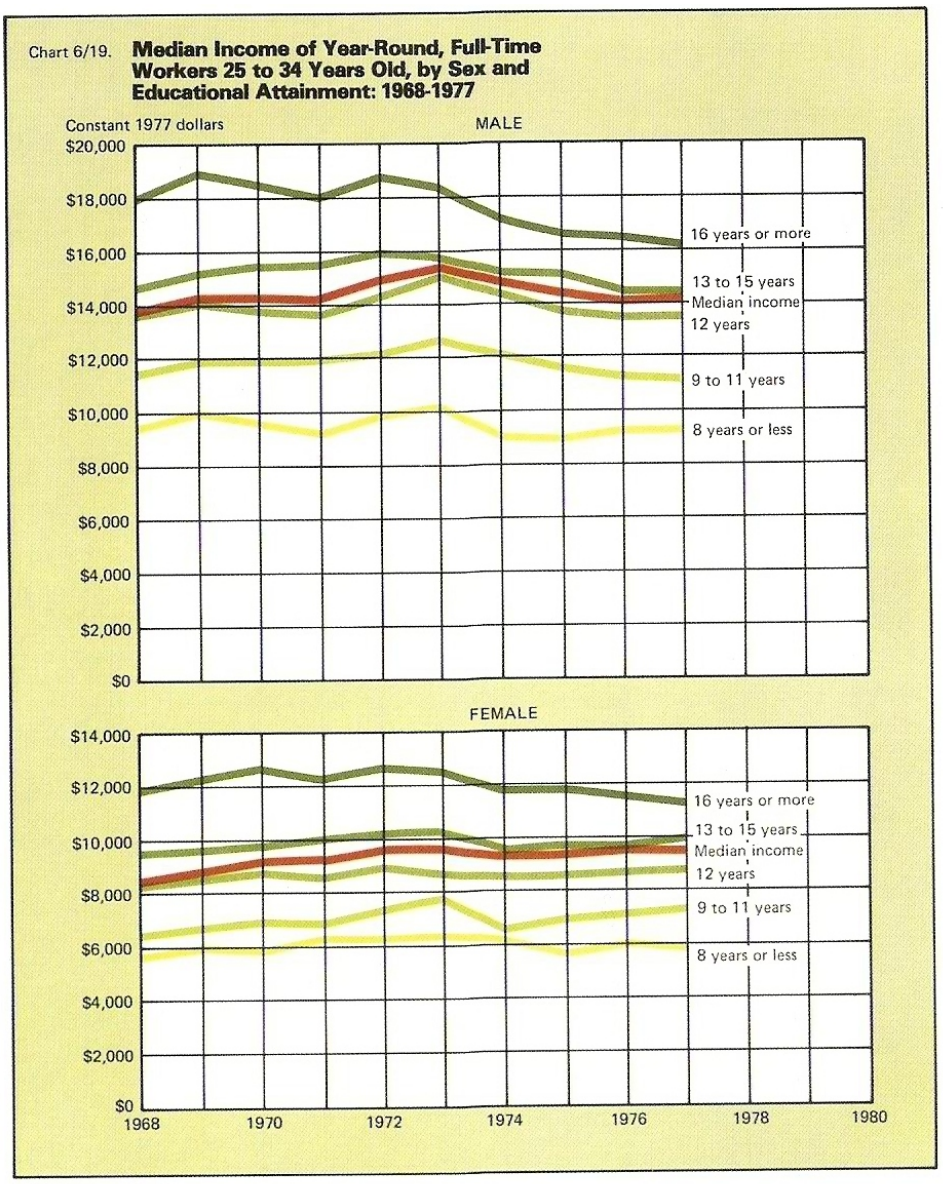


Figure 21: Wainer (1997), p. 30, Figure 26: Wainer (1997), p. 29, Figure 25, improved.

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.

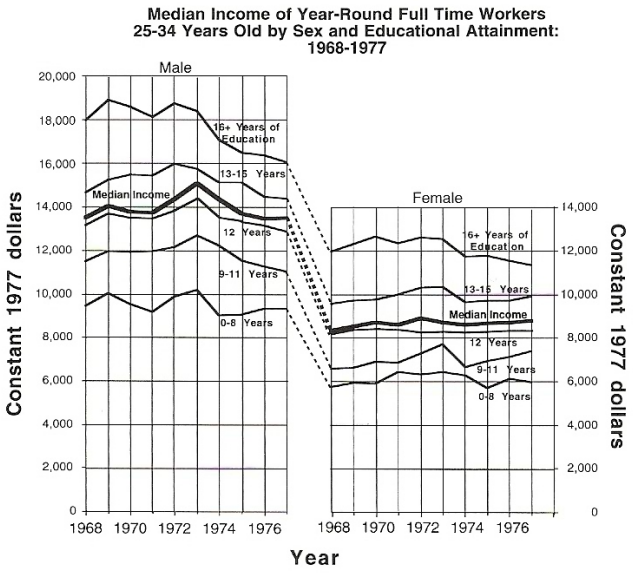


FIGURE 28. Figure 27 redone with the two plots horizontally opposed, showing the size of sex differences more clearly.

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

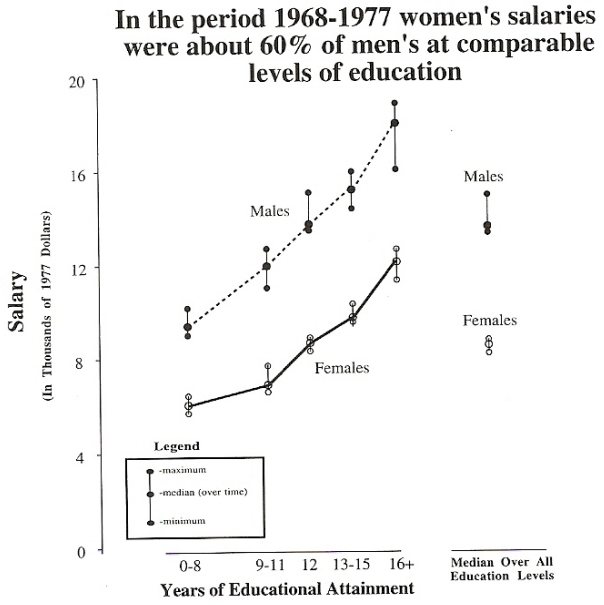
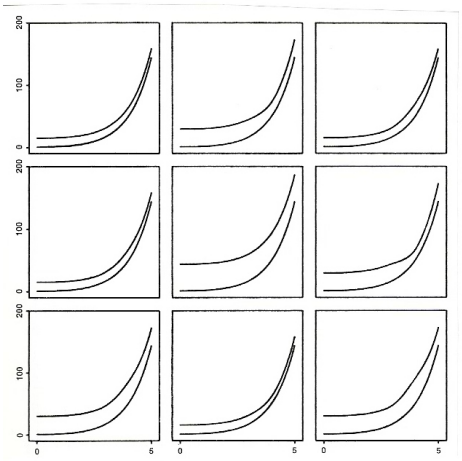


FIGURE 29. Figure 28 redone with the large effects of sex and education emphasized and the small-time trend suppressed.

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.

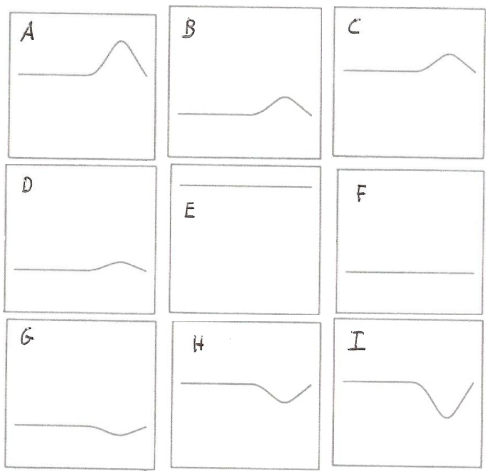
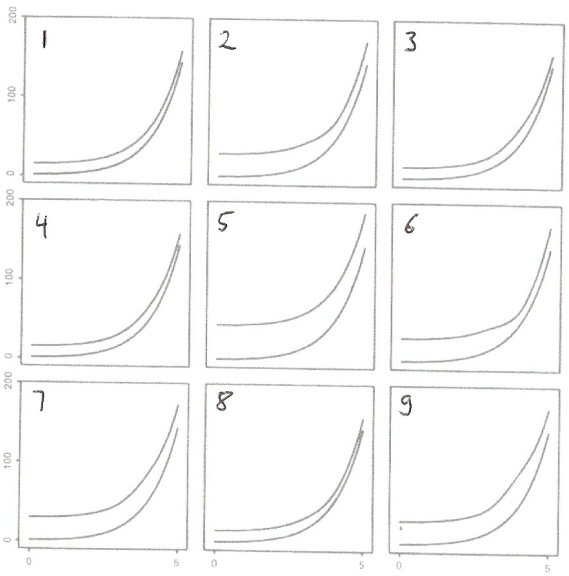
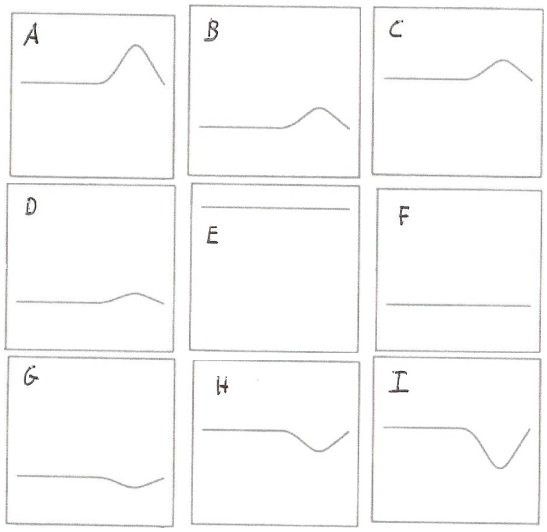


Figure 26: Wainer (1997), p. 33, Figure 30: Scrambled differences. The horizontal axis covers the interval 0 to 5, the vertical axis covers the interval 0 to 50.

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: _____

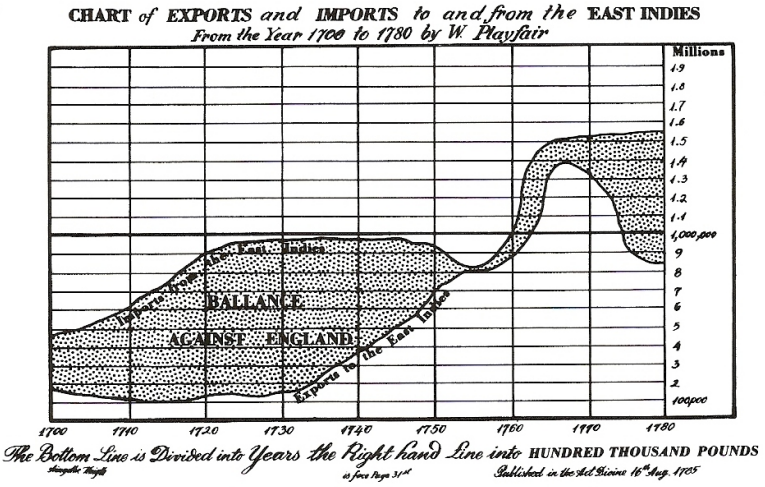


FIGURE 31. William Playfair's eighteenth-century graph of England's imports and exports with the East Indies (from Cleveland and McGill, 1984).

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

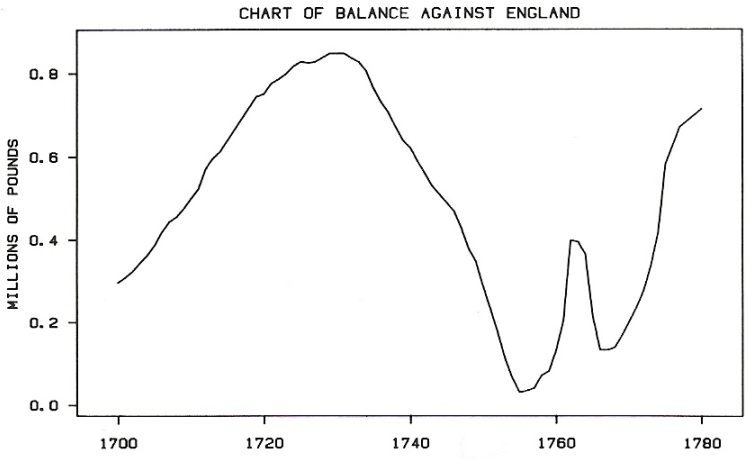


FIGURE 32. A graph of the difference between West Indies imports and exports showing explicitly the previously invisible jump in the 1760s (from Cleveland and McGill, 1984).

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

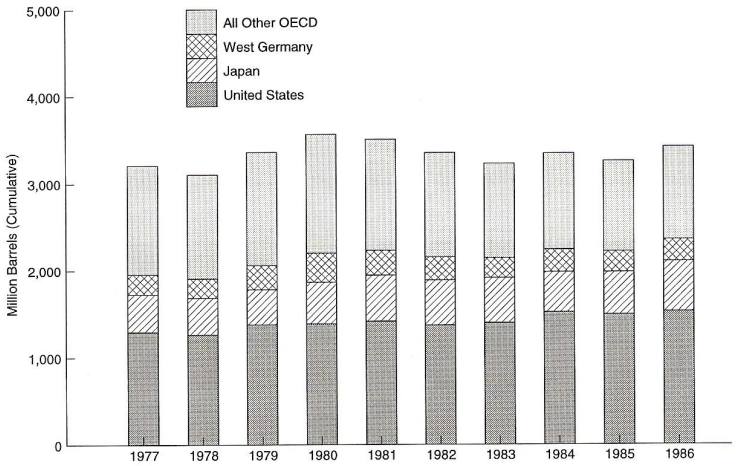


FIGURE 33. From the U.S. Department of Energy's *Annual Energy Review, 1986,* showing the changes in primary stocks of petroleum in OECD countries.

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

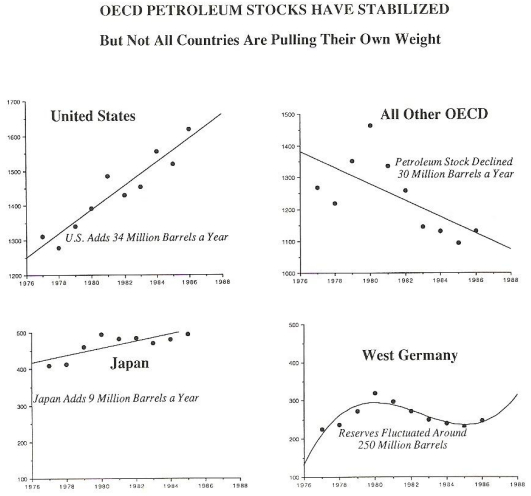
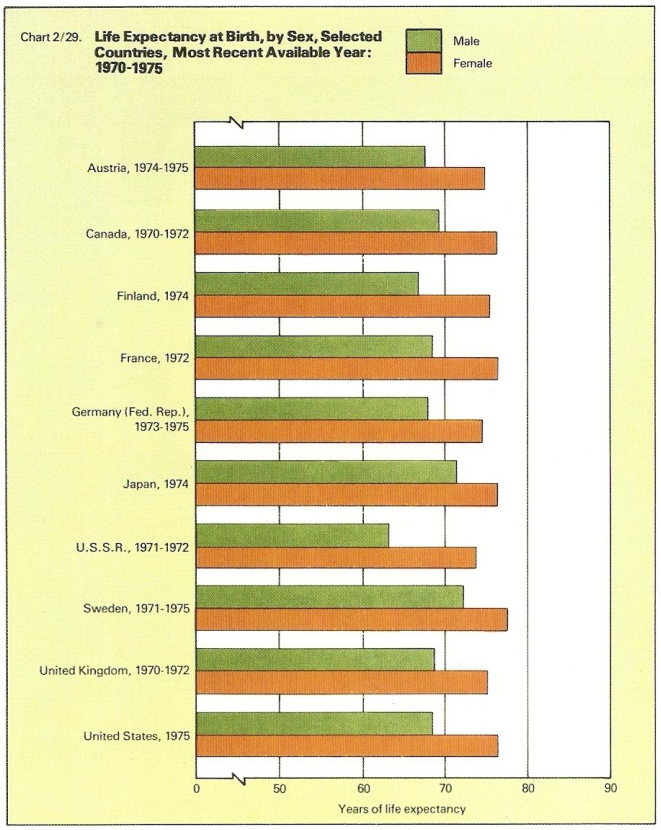


FIGURE 34. Regraphing of the data from figure 33 in which each country's data are shown relative to a straight line.

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.

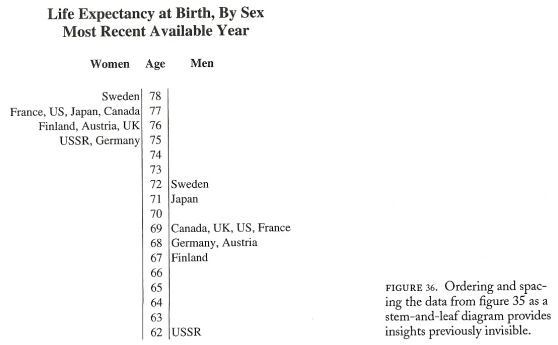
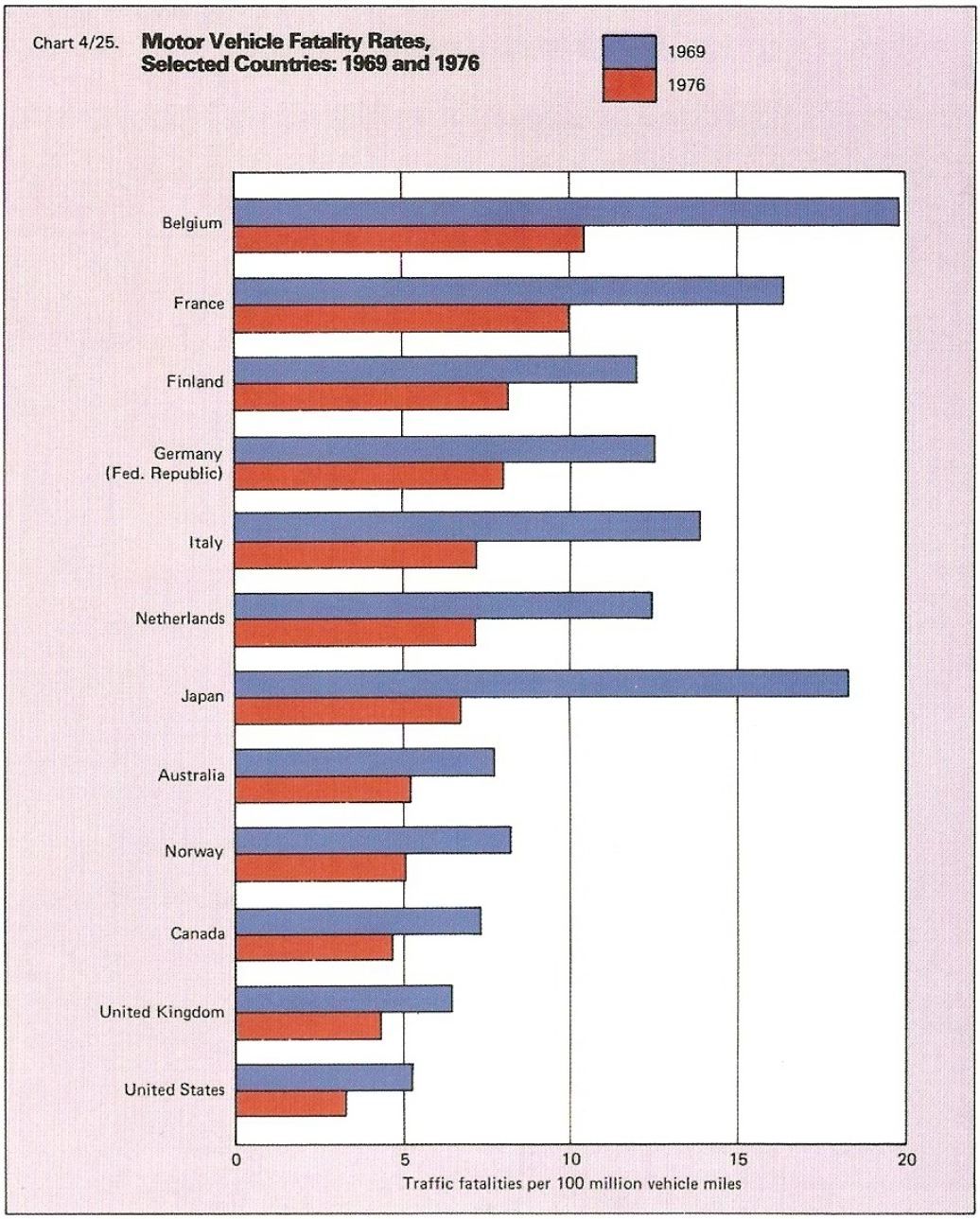


FIGURE 36. Ordering and spacing the data from figure 35 as a stem-and-leaf diagram provides insights previously invisible.

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.

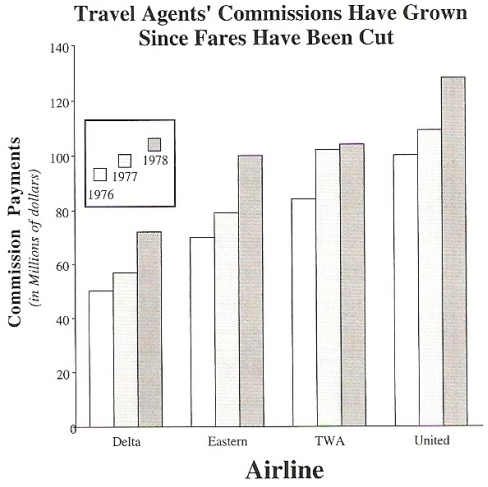


FIGURE 39. Figure 38 redrawn with 1978 data placed on a comparable basis shows that the fare cuts have been a boon to travel agents.

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.

TABLE 1
Life Expectancy at Birth

Country	Male	Female
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.

TABLE 2
Life Expectancy at Birth

Country	Male	Female
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.

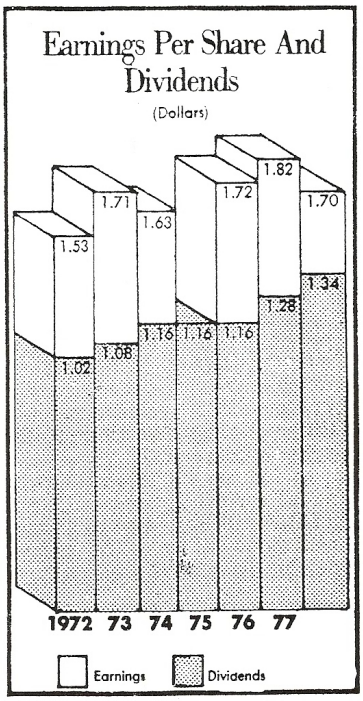


FIGURE 41. An extra dimension on earnings and dividends confuses even the grapher (from the *Washington Post*, 1979).

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

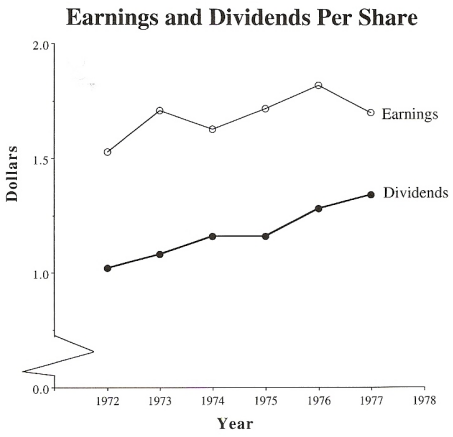
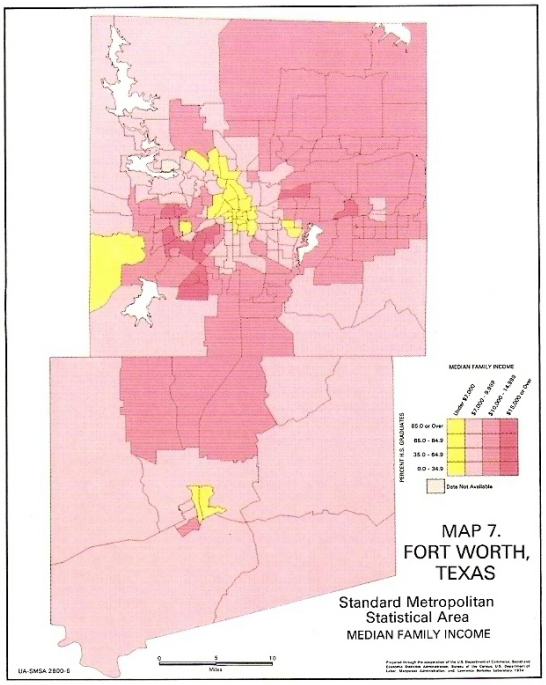


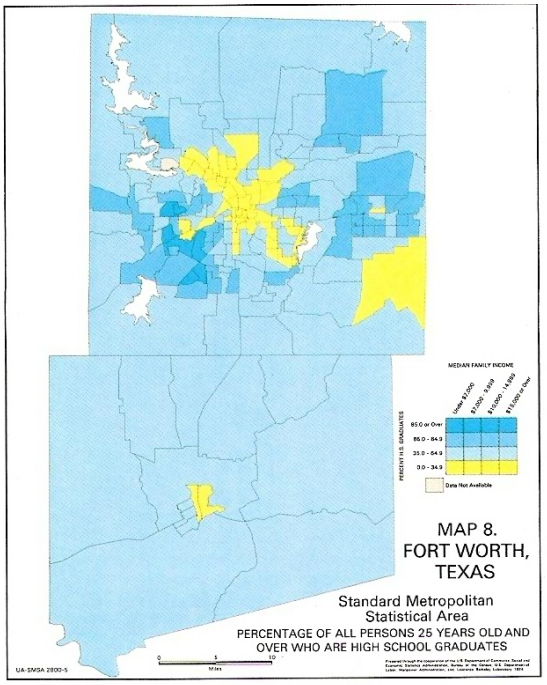
FIGURE 42. Data from figure 41 redrawn simply.

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.

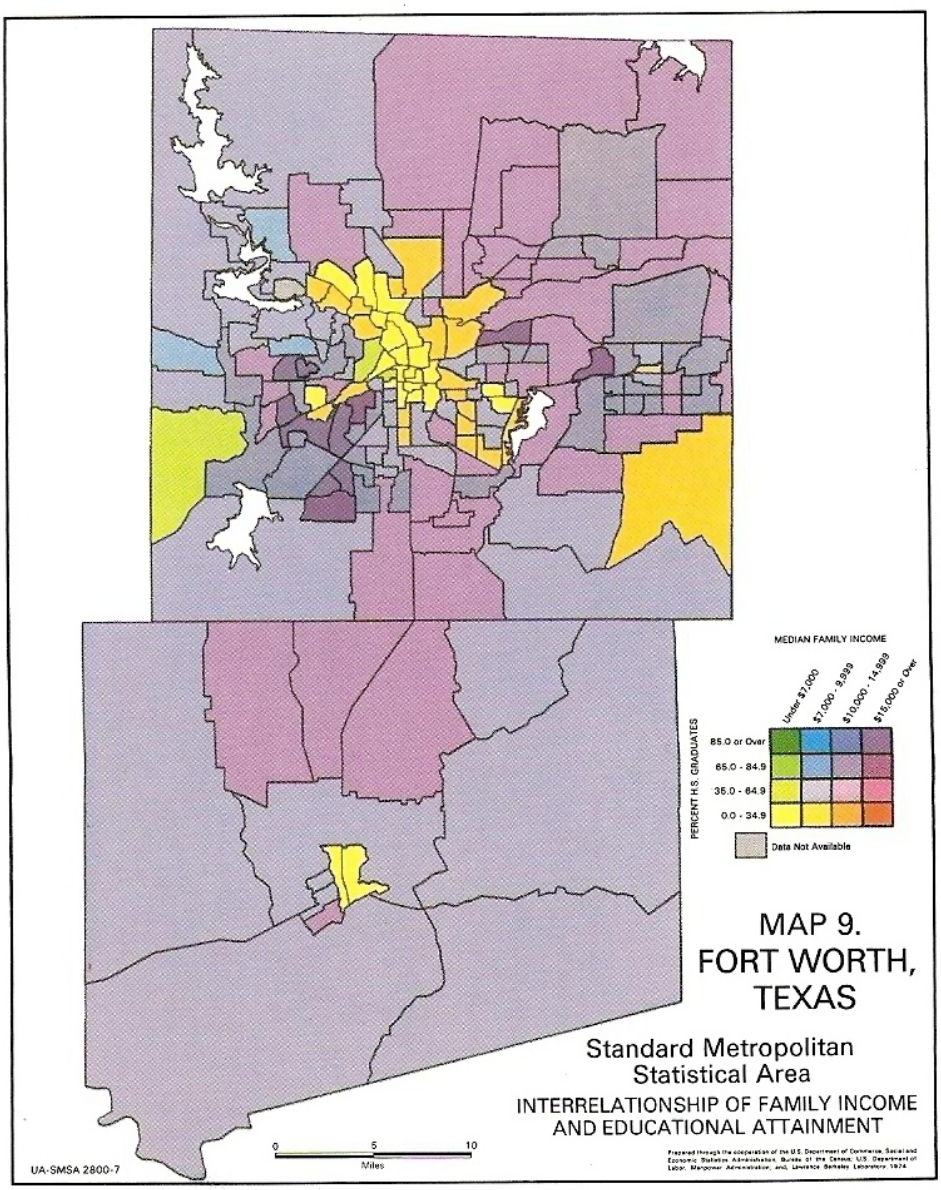


CHAPTER 1, FIGURE 44. The geographic distribution of median family income in Fort Worth, Texas, in 1974.



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

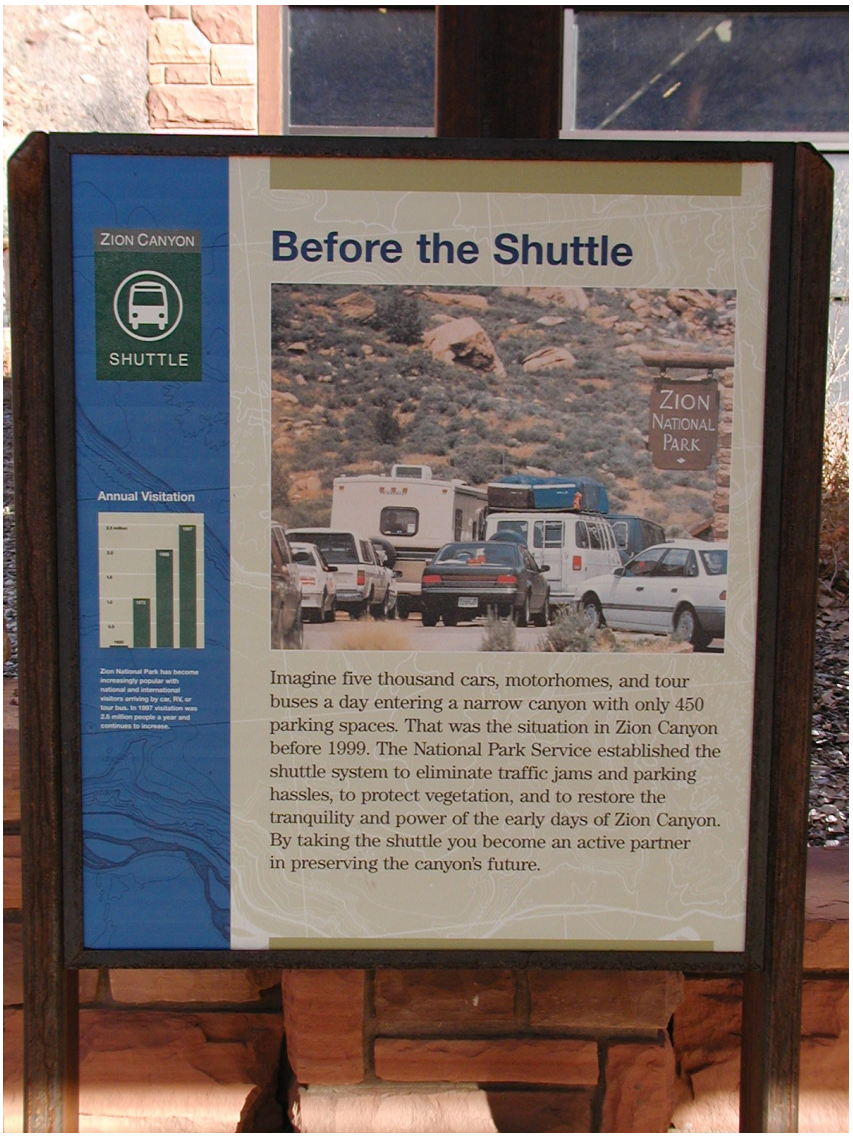


Figure 42: Personal Photograph: From the distance, the annual visitation appears to increase linearly, ...

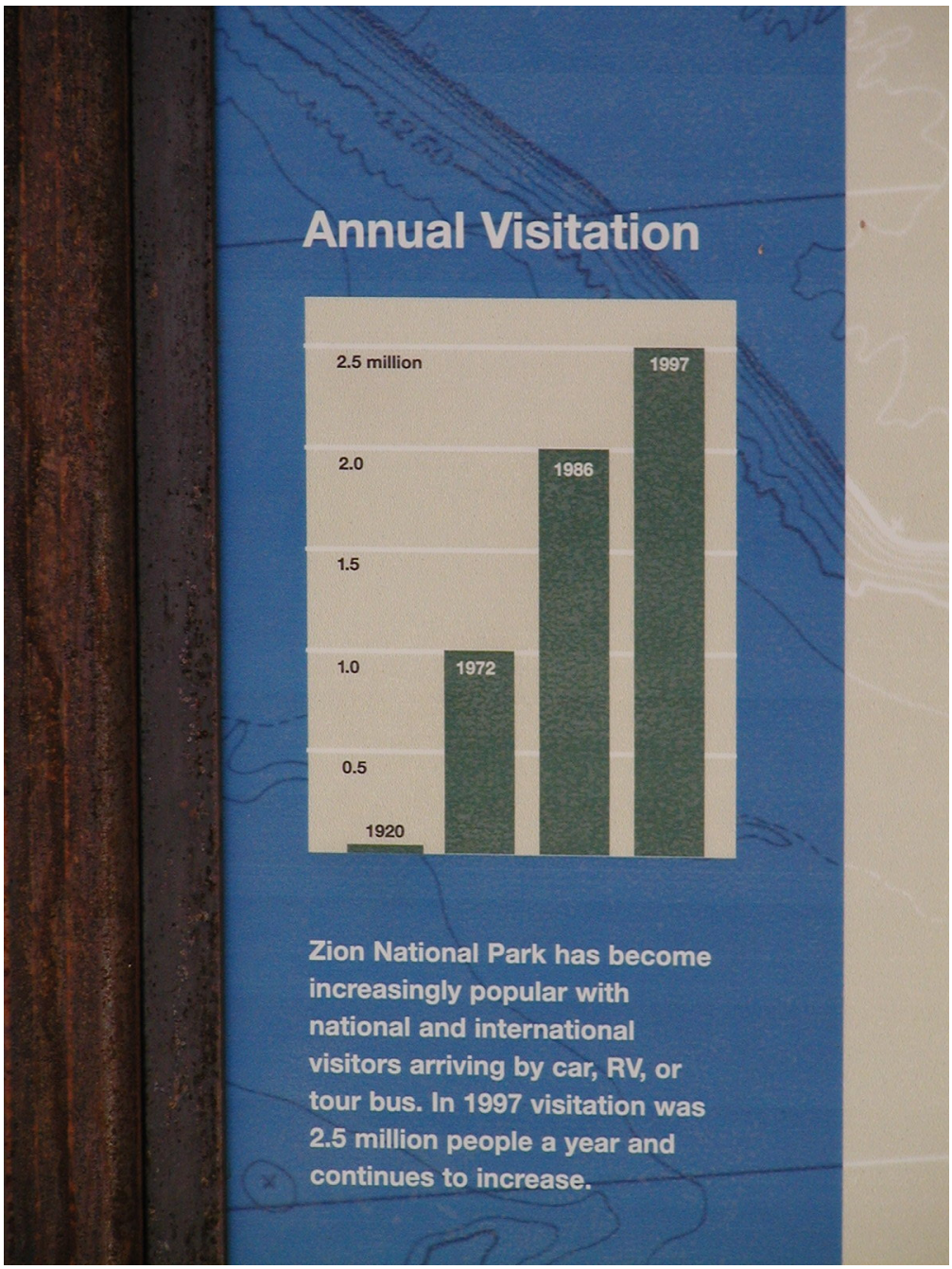


Figure 43: Personal Photograph: . . . but at a closer view, this is certainly not the case.

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.

Improved Version:

`http://www.math.usu.edu/~symanzik/teaching/2009_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/2009_stat6560/RDataAndScripts/Berlin.R`

Example 3: Wikipedia, 2009

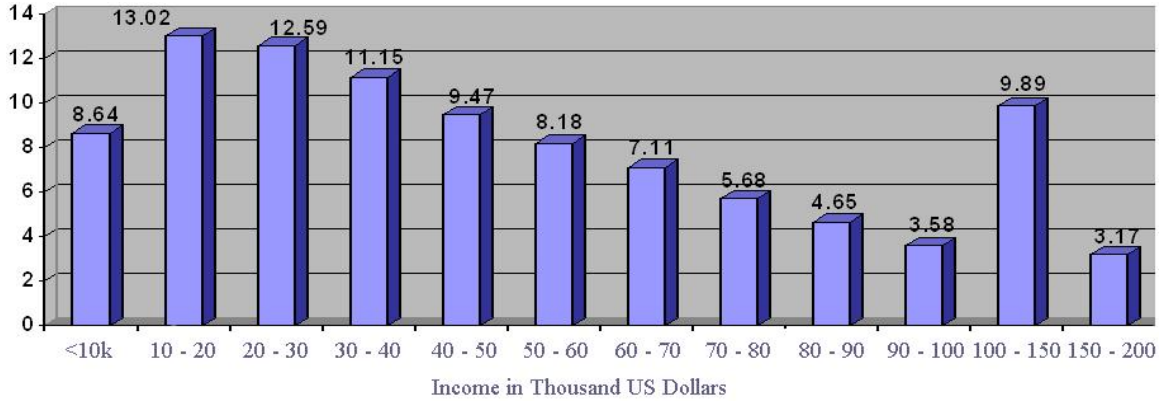


Figure 46: Figure taken from `http://en.wikipedia.org/wiki/Household_income_in_the_United_States` on 1/13/2009.

Rules followed (to make this a bad graphic):

- 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/2009_stat6560/RDataAndScripts/Wiki.R`

Example 4: Computational Statistics, 2002

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).

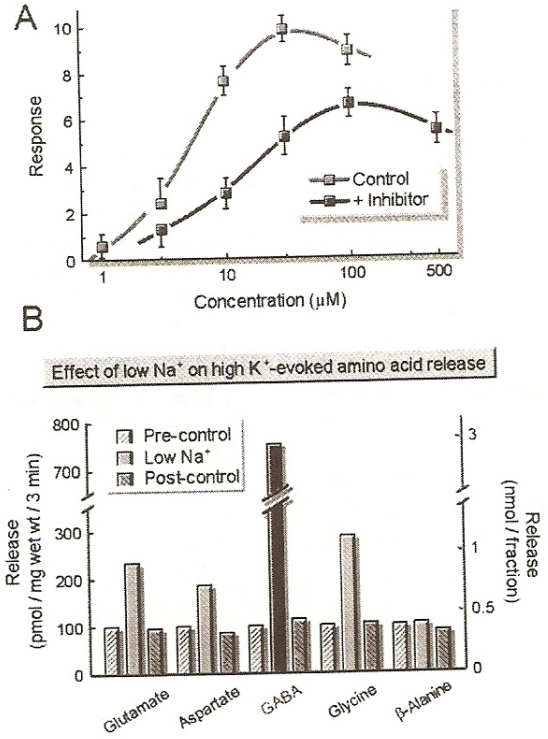


Figure 3: Line and bar graphs created with KyPlot

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 signifcant 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 Na^+ differs considerably, though, for the various amino acids.

Improved Version:

`http://www.math.usu.edu/~symanzik/teaching/2009_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?
- 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 consistenly?

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)

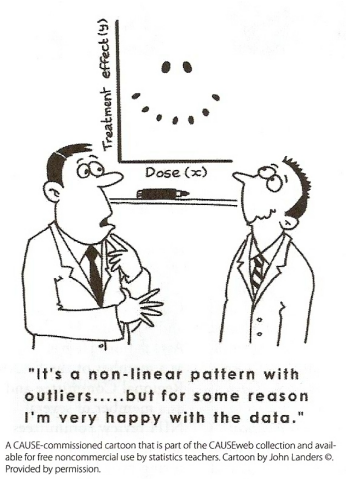


Figure 48: Amstat News, January 2009, p. 25, Cartoon.

Lecture 11: Fr 01/30/09

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 (www.infografe.com.br). `http://www.math.yorku.ca/SCS/Gallery/milestone/Visualization_Milestones.pdf` (original) `http://www.math.usu.edu/~symanzik/teaching/2009_stat6560/Downloads/Friendly_Visualization_Milestones.pdf` (local copy obtained on 1/25/2009)
- *"Milestones in the History of Thematic Cartography, Statistical Graphics, and Data Visualization"*. Michael Friendly, October 16, 2008. `http://www.math.yorku.ca/SCS/Gallery/milestone/milestone.pdf` (original) `http://www.math.usu.edu/~symanzik/teaching/2009_stat6560/Downloads/Friendly_milestone.pdf` (local copy obtained on 1/25/2009)

Lecture 12:
Mo 02/02/09

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-Ű1899: Golden Age of Data Graphics

1900Ű-1949: Modern Dark Ages

1950Ű-1974: Re–birth of Data Visualization

1975Ű-present: High–D Data Visualization

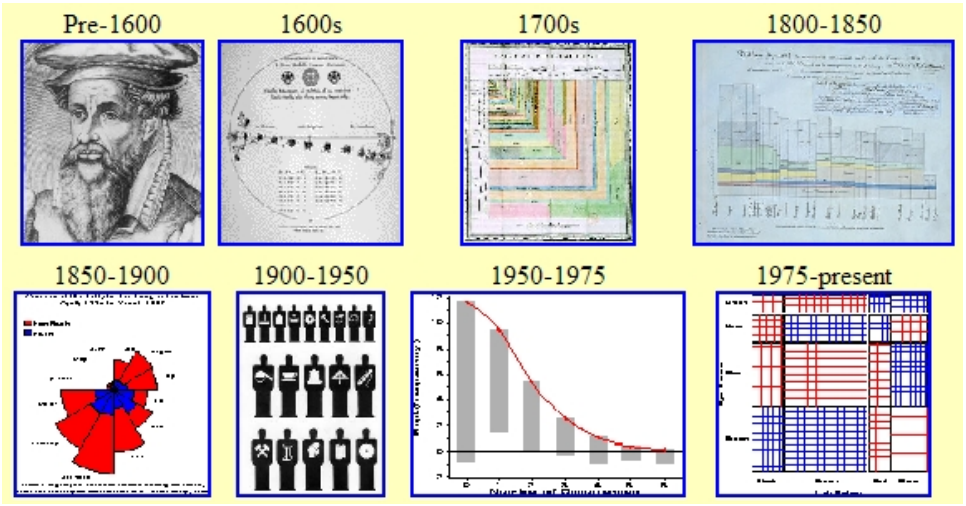


Figure 49: Screenshot taken from `http://www.math.yorku.ca/SCS/Gallery/milestone/` on 1/25/2009.

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: (1629Ű-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.

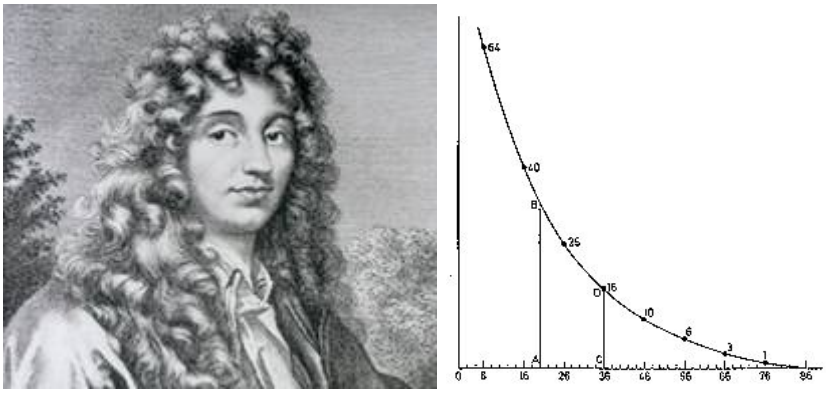


Figure 50: Figures taken from `http://www.math.yorku.ca/SCS/Gallery/images/portraits/huygens.jpg` and `http://www.math.yorku.ca/SCS/Gallery/images/huygens-graph.gif` on 1/27/2009.

William Playfair: (1759Ű-1823), England

1786: Bar chart, line graphs of economic data.

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

See Project 1.3 for more details.

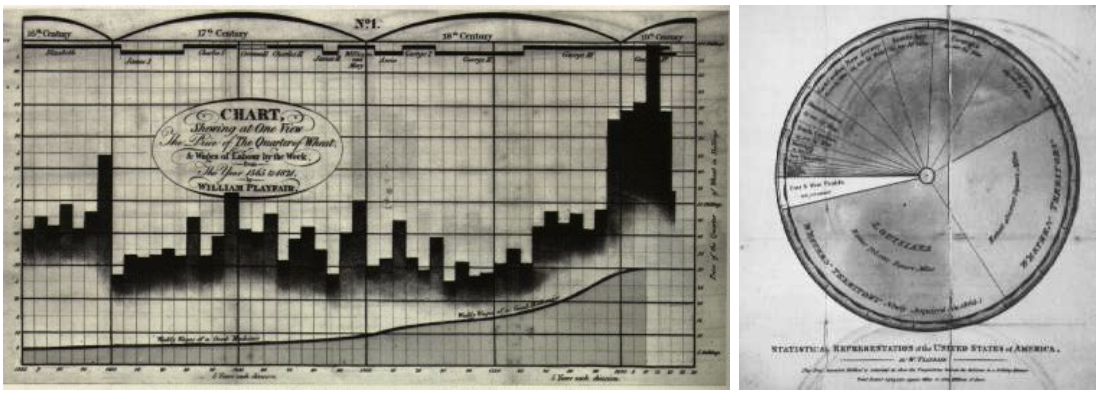


Figure 51: Figures taken from `http://www.math.yorku.ca/SCS/Gallery/images/playfair-wheat1.gif` and `http://www.math.yorku.ca/SCS/Gallery/images/playfair-pie.jpg` on 1/27/2009.

Charles Joseph Minard: (1781Ű-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 March on Moscow.

See Project 1.4 for more details.

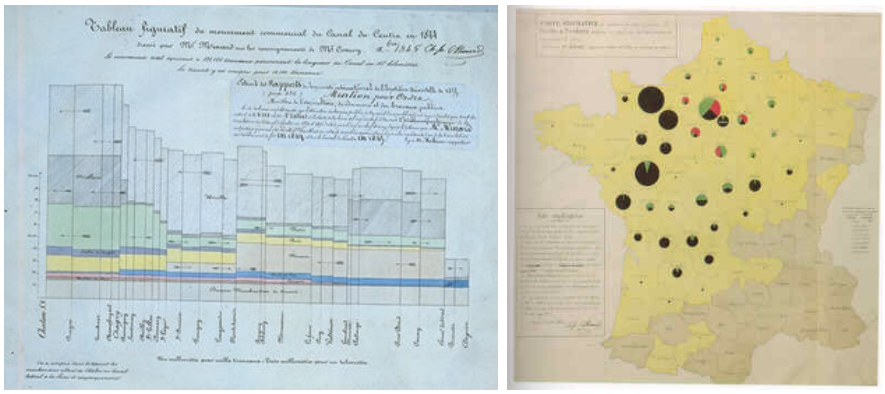


Figure 52: Figures taken from `http://www.math.yorku.ca/SCS/Gallery/images/enpc/img09a.jpg` and `http://www.math.yorku.ca/SCS/Gallery/images/Robinson/viandes.jpg` on 2/1/2009.

Florence Nightingale: (1820Ű-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`.

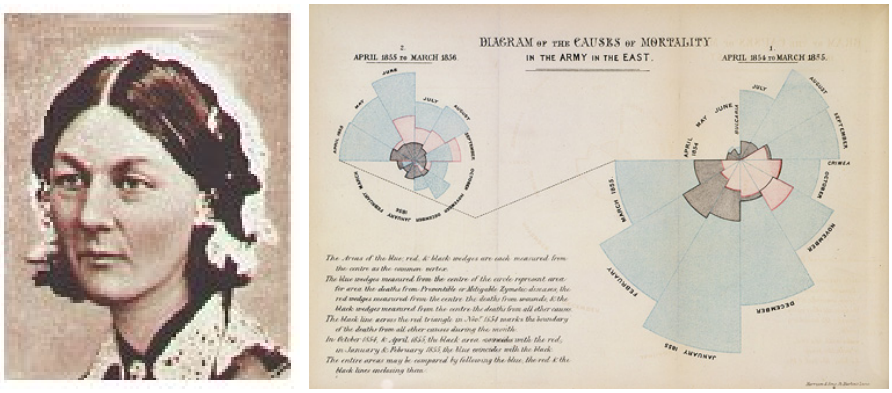


Figure 53: Figures taken from `http://www.math.yorku.ca/SCS/Gallery/images/portraits/nightingale.jpg` and `http://en.wikipedia.org/wiki/File:Nightingale-mortality.jpg` on 1/27/2009.

Francis Galton: (1822–1911), England

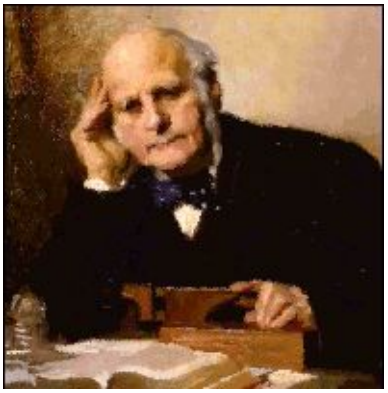
1861: The modern weather map, a chart showing area of similar air pressure and barometric changes by means of glyphs displayed on a map. These led to the discovery of the anti–cyclonic movement of wind around low–pressure areas.

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.



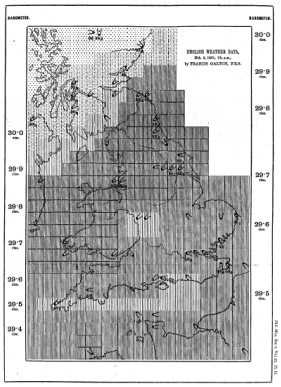


Figure 54: Figures taken from `http://www.math.yorku.ca/SCS/Gallery/images/portraits/galton-furse.gif` and `http://galton.org/essays/1860-1869/galton-1861-charts.pdf` on 1/27/2009.

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 "scatterplot 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 Ů– Paul A. Tukey and John W. Tukey.

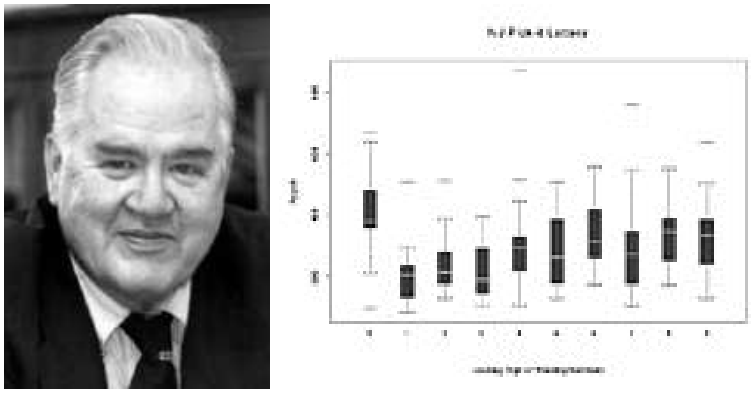


Figure 55: Figures taken from `http://www.math.yorku.ca/SCS/Gallery/images/portraits/tukey2.jpg` and `http://www.math.yorku.ca/SCS/Gallery/icons/NJPick-it.gif` on 2/1/2009.

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.

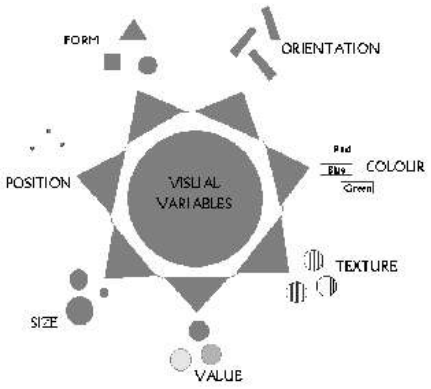


Figure 56: Figures taken from `http://www.math.yorku.ca/SCS/Gallery/images/portraits/jbertin.jpg` and `http://www.math.yorku.ca/SCS/Gallery/images/bertin-ve.jpg` on 2/1/2009.

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.

See Project 1.6 for more details.

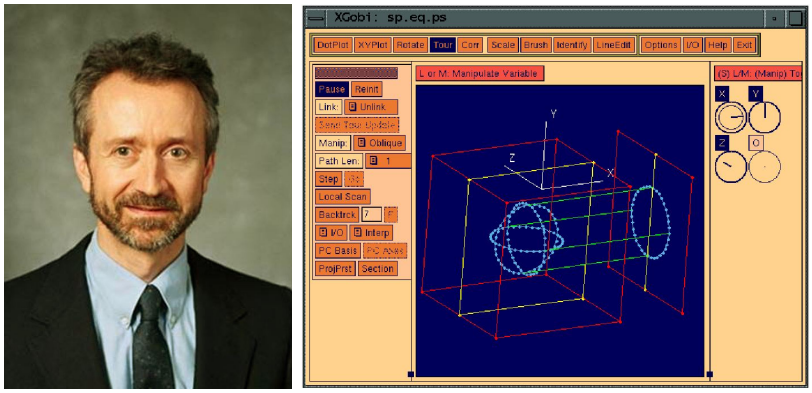
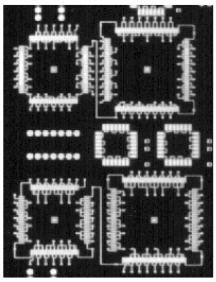


Figure 57: Figures taken from `http://www.math.yorku.ca/SCS/Gallery/people/AndreasBuja.jpg` and `http://www.research.att.com/areas/stat/xgobi/` on 2/1/2009.

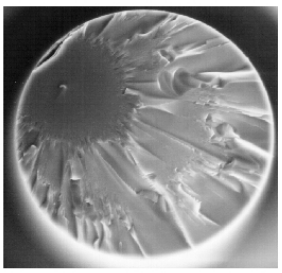
John W. Chambers and Collaborators: USA

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 58: 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.

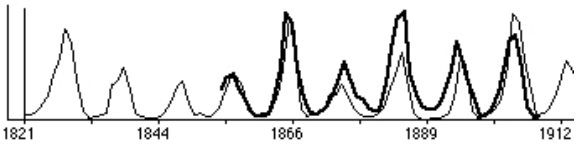


Figure 59: Figures taken from `http://stats.math.uni-augsburg.de/~unwin/` and `http://www.math.yorku.ca/SCS/Gallery/images/DiamondFast.jpg` on 2/1/2009.

Edward J. Wegman: USA

Personal Home Page: `http://www.galaxy.gmu.edu/stats/faculty/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.

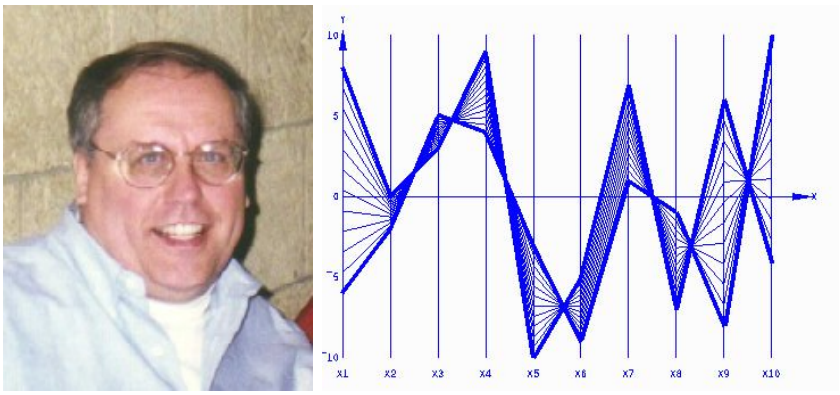


Figure 60: Figures taken from `http://www.math.yorku.ca/SCS/Gallery/people/EdWegman.jpg` and `http://www.math.yorku.ca/SCS/Gallery/images/parallel-coords.gif` on 2/1/2009.

2.3 Statistical Graphics and Events in History

Project 1.1: "John Snow and the Cholera Epidemic in London, 1854."

Project 1.2: "The Challenger Disaster in 1986: How Graphics played a Deadly Role."

2.4 Further Reading

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

- Brillinger (2002) (preprint available at `http://www.stat.berkeley.edu/~brill/Papers/life.pdf`)
- 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) (see also Homework 1, Question (iv))

Lecture 13:
We 02/04/09

3 Use of Color

3.1 Color–Deficiency and Color–Blindness

Tufte (1983), p. 183, states:

> "There are many specific differences between friendly and unfriendly graphics. [...]
>
> Friendly: colors, if used, are chosen so that the color–deficient and color–blind (5 to 10 percent of viewers) can make sense of the graphic (blue can be distinguished from other colors by most color–deficient people). [...]
>
> Unfriendly: design insensitive to color–deficient viewers; red and green used for essential contrasts. [...]"

Question: Which numbers and letters can you see hidden in each figure?

Figure 61: Illustration of various tests for color blindness. Figure taken from `http://www.nlm.nih.gov/medlineplus/colorblindness.html` on 2/3/2009.

Definitions:
"**Color blindness**, a color vision deficiency, is the inability to perceive differences between some of the colors that others can distinguish. It is most often of genetic nature, but may also occur because of eye, nerve, or brain damage, or due to exposure to certain chemicals. The English chemist John Dalton published the first scientific paper on the subject in 1798, "Extraordinary facts relating to the vision of colours", after the realization of his own color blindness; because of Dalton's work, the condition is sometimes called Daltonism, although this term is now used for a type of color blindness called deuteranopia. [...]

Dichromacy: Protanopes, deuteranopes, and tritanopes are dichromats; that is, they can match any color they see with some mixture of just two spectral lights (whereas normally humans are trichromats and require three lights). These individuals normally know they have a color vision problem and it can affect their lives on a daily basis. Protanopes and deuteranopes see no perceptible difference between red, orange, yellow, and green. All these colors that seem so different to the normal viewer appear to be the same color for this two percent of the population. [...]

Protanopia (1% of males): Lacking the long–wavelength sensitive retinal cones, those with this condition are unable to distinguish between colors in the green–yellow–red section of the spectrum. [...]

Deuteranopia (1% of males): Lacking the medium–wavelength cones, those affected are again unable to distinguish between colors in the green–yellow-red section of the spectrum. [...]

Tritanopia (less than 1% of males and females): Lacking the short–wavelength cones, those affected are unable to distinguish between the colors in the blue-yellow section of the spectrum. [...]

Anomalous trichromacy: Those with protanomaly, deuteranomaly, or tritanomaly are trichromats, but the color matches they make differ from the normal. They are called anomalous trichromats. [...]

Protanomaly (1% of males, 0.01% of females): Having a mutated form of the long–wavelength (red) pigment, whose peak sensitivity is at a shorter wavelength than in the normal retina, protanomalous individuals are less sensitive to red light than normal.[...] This causes reds to reduce in intensity to the point where they can be mistaken for black. [...]

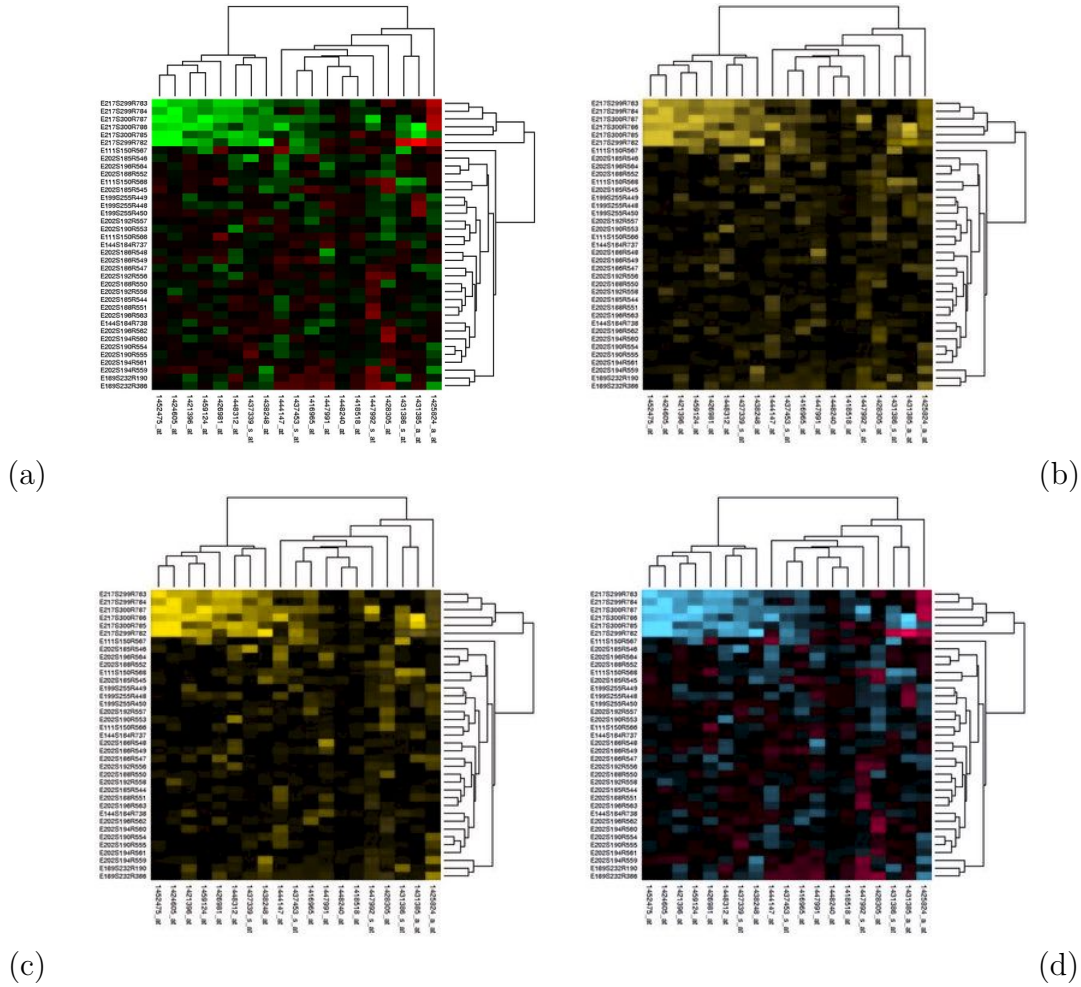
Deuteranomaly (most common — 6% of males, 0.4% of females): Having a mutated form of the medium–wavelength (green) pigment. [...] This is the most common form of color blindness, making up about 6% of the male population. The deuteranomalous person is considered "green weak". For example, in the evening, dark green cars appear to be black to Deuteranomalous people. Similar to the protanomates, deuteranomates are poor at discriminating small differences in hues in the red, orange, yellow, green region of the spectrum.[...]

Tritanomaly (equally rare for males and females): Having a mutated form of the short–wavelength (blue) pigment. [...]"

(Definitions taken from `http://en.wikipedia.org/wiki/Color_blindness` on 2/3/2009.)

See `http://www.nlm.nih.gov/medlineplus/colorblindness.html` for additional information and links.

Question: How does a figure appear for someone with a particular color vision deficiency? You can upload a figure of interest to `http://www.vischeck.com/vischeck/vischeckImage.php` and then obtain the simulated result.



(a) (b) (c) (d)

Figure 62: Red–green heatmap (a) taken from `http://en.wikipedia.org/wiki/Image:Heatmap.png` on 2/3/2009. Simulation of three main types of color vision deficiencies: (b) Deuteranope (a form of red/green color deficit); (c) Protanope (another form of red/green color deficit); and (d) Tritanope (a blue/yellow deficit — very rare), obtained from `http://www.vischeck.com/vischeck/vischeckImage.php` on 2/3/2009.

Task: Start with the R code at `http://www.math.usu.edu/~symanzik/teaching/2009_stat6560/RDataAndScripts/Ch3_TestColors.R` and modify the colors. Save your figure as a jpg and run it through the different simulations for color vision deficiencies accessible at `http://www.vischeck.com/vischeck/vischeckImage.php`. After having tested a few options, can you suggest what may work well for all viewers (and what may not work well for some viewers)?

3.2 Various Color Spaces

(Based on Kosslyn (1994), Chapter 7, Kosslyn (2006), Chapter 7 & Few (2004), Chapter 2)

3.2.1 The HSL and HSV Color Spaces

Color is not a single entity, but it can be broken down into three components:

Hue (H): This is what we usually mean by color. It is the qualitative aspect which depends on the wavelength of the light (from long at the red end of the spectrum to short at the violet end).

Saturation (S): This is the deepness of the color (which can be varied by the amount of white that is added).

Lightness (L), Value (V), Intensity (I), or Brightness (B): This is the amount of light that is reflected (if shown on a printed page) or that is emitted (if the display is projected from a slide or shown on a computer screen). In either case, intensity can be varied by the amount of gray that is added.

Color encoded via hue, saturation, and lightness is called a HSL color space. Equivalent color spaces are based on hue, saturation, and intensity (HSI) or on hue, saturation, and brightness (HSB). Still using the idea of adding gray as the third component, but using a different encoding results in the hue, saturation, and value (HSV) color space. Sometimes, the last two letters of the color space are swapped, so we may speak of a HLS (instead of a HSL) or a HVS (instead of HSV) color space.

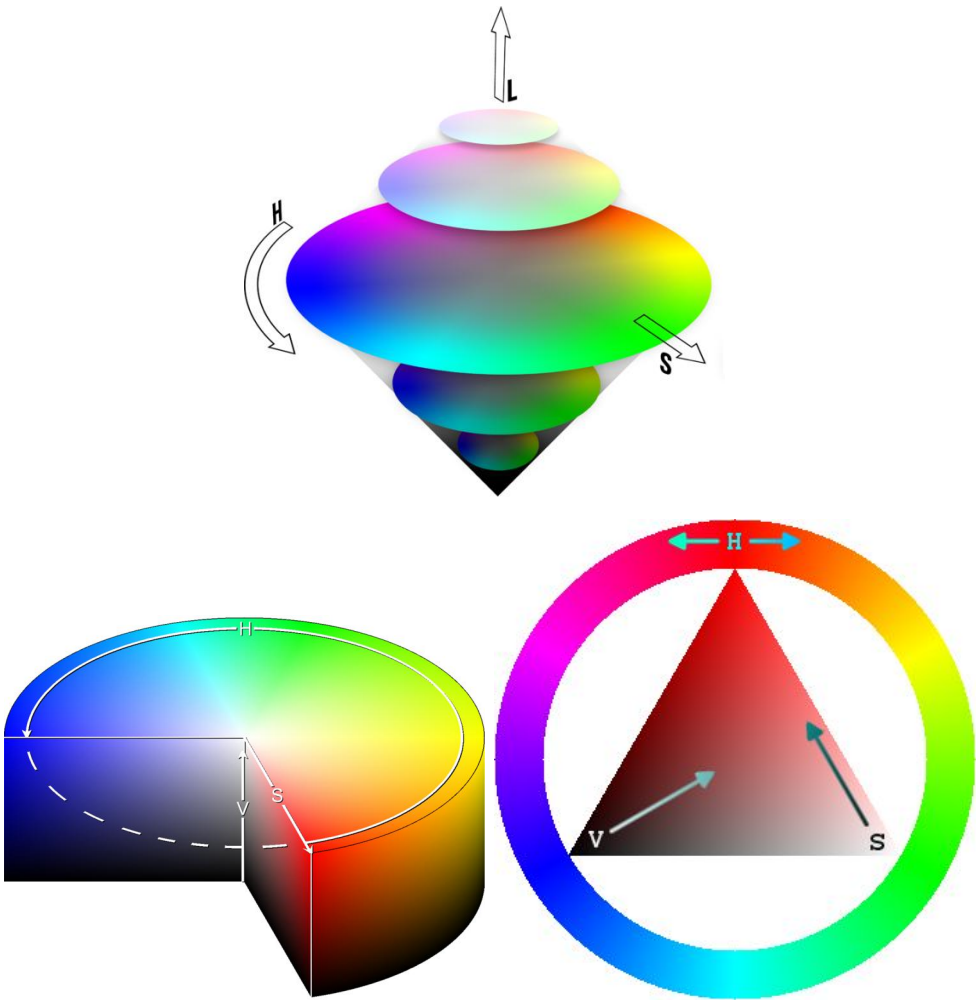


Figure 63: Illustration of the HSL (top) and HSV (bottom) color spaces. Figures taken from `http://en.wikipedia.org/wiki/File:Color_cones.png`, `http://en.wikipedia.org/wiki/File:HSV_cylinder.png` and `http://en.wikipedia.org/wiki/File:Triangulo_HSV.png` on 2/5/2009 (top) and 2/3/2009 (bottom).

3.2.2 The RGB Color Space

The **RGB color model** is an additive color model in which red, green, and blue light are added together in various ways to reproduce a broad array of colors. The name of the model comes from the initials of the three additive primary colors, red (R), green (G), and blue (B).
(Definition taken from `http://en.wikipedia.org/wiki/RGB_color_model` on 2/3/2009.)

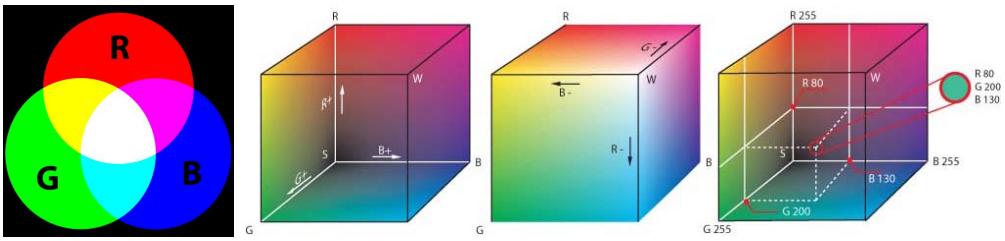


Figure 64: Illustration of the RGB color model and color space. Figures taken from `http://en.wikipedia.org/wiki/File:AdditiveColor.svg` and `http://en.wikipedia.org/wiki/File:RGB_farbwuerfel.jpg` on 2/3/2009.

In many programming languages including R, the RGB color space is being used as the primary (and easiest to use) color space. Values for each of the three components (R, G, and B) can originate from the discrete set $\{0, \ldots, 255\}$ or the continuous interval $[0 \ldots 1]$.

Some main color combinations in RGB are:

Color name	Red	Green	Blue	Hexadecimal
Black	0	0	0	#000000
White	255	255	255	#FFFFFF
Red	255	0	0	#FF0000
Green	0	255	0	#00FF00
Blue	0	0	255	#0000FF
Yellow	255	255	0	#FFFF00

art with the R code at `http://www.math.usu.edu/~symanzik/teaching/2009_stat6560/RDataAndScripts/Ch3_RGBColors.R` where we vary the red (R) component in a linear

way. Can we clearly distinguish among the 10 different colors? If not, about how many visually distinct colors can you easily identify?

Repeat the same for the green (G) and the blue (B) component. How many visually distinct colors can you easily identify for these?

In contrast, repeat the same with a variation of gray levels where each gray level can be obtained by tripling the same value x, i.e., (x, x, x). Is this better? Let's see whether we can obtain even some further improvement later on ...

3.2.3 The HCL Color Space

This color space consists of the following three components:

Hue (H): As before, this component describes the dominant wavelength.

Chroma (C): This component describes the colorfulness, i.e., the intensity of the color as compared to gray.

Luminance (L): This component relates to brightness, i.e., the amount of gray.

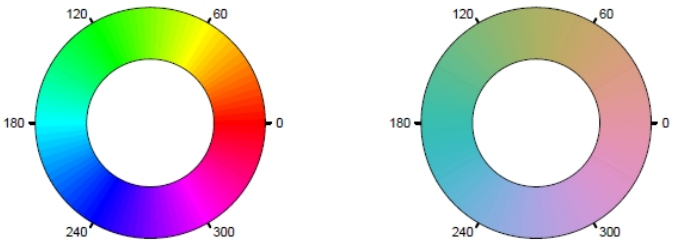


Figure 3: HSV-based and HCL-based color wheel.

Figure 65: Zeileis et al. (2008), p. 6, Figure 3.

3.3 Suggestions for Color Selections

(Based on Kosslyn (1994), Chapter 7 & Kosslyn (2006), Chapter 7)

Similar to our lists in Chapter 1 how to create bad and good graphics, there exist suggestions how to use color in graphics. Below the suggestions from Kosslyn (2006), Chapter 7:

- Use colors that are well separated in the spectrum.
- Make adjacent colors have different brightness.

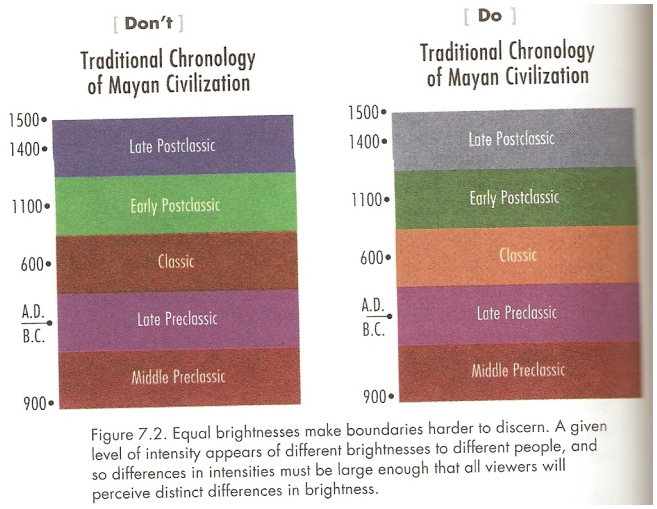


Figure 66: Kosslyn (2006), p. 160, Figure 7.2.

- Make the most important content element the most salient.

- Use warm colors to define a foreground.

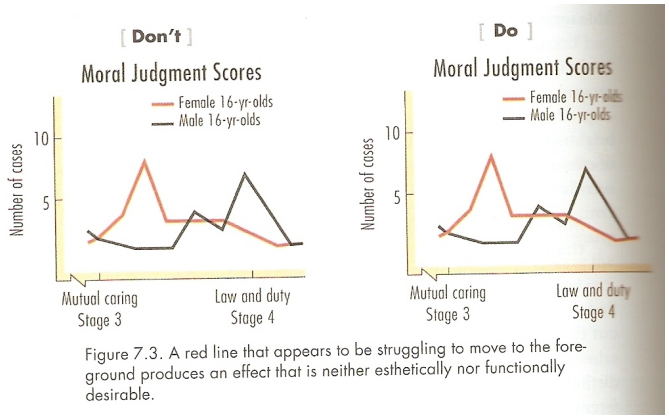


Figure 7.3. A red line that appears to be struggling to move to the foreground produces an effect that is neither esthetically nor functionally desirable.

Figure 67: Kosslyn (2006), p. 162, Figure 7.3.

- Avoid using red and blue in adjacent regions.

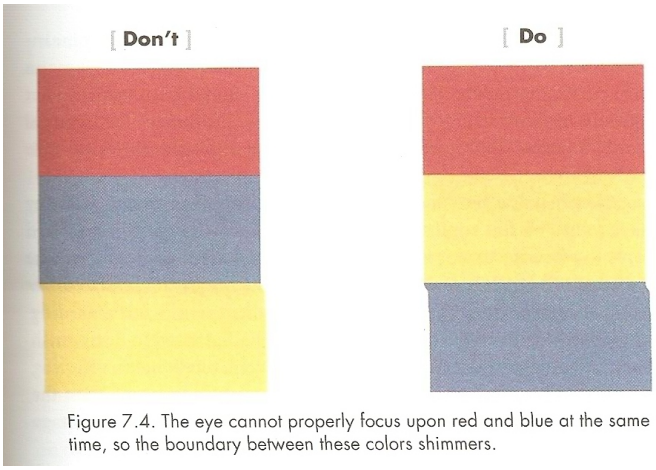


Figure 7.4. The eye cannot properly focus upon red and blue at the same time, so the boundary between these colors shimmers.

Figure 68: Kosslyn (2006), p. 163, Figure 7.4.

- Respect compatibility and conventions of color.

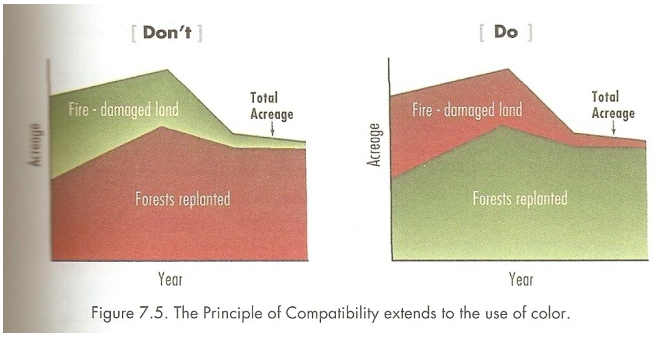


Figure 69: Kosslyn (2006), p. 163, Figure 7.5.

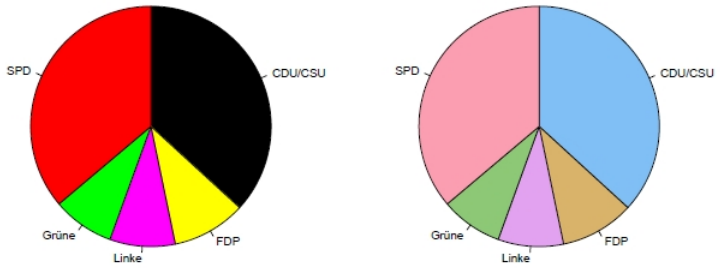


Figure 70: Zeileis et al. (2008), p. 12, Figure 10: While HCL–based color palattes are a worthwhile alternative to standard RGB color choices, the right part of this figure clearly should be labeled "**[Don't]**. The original colors (black, red, yellow, green, and purple) are so strongly associated with the German parties that they shouldn't be changed. Imagine that someone would change the tradional red–blue US election map to pink–cyan, or, even worse, that someone would change the colors of traffic lights from red–yellow–green to something different because that is more appealing to the viewer — unthinkable.

- Use color to group elements.

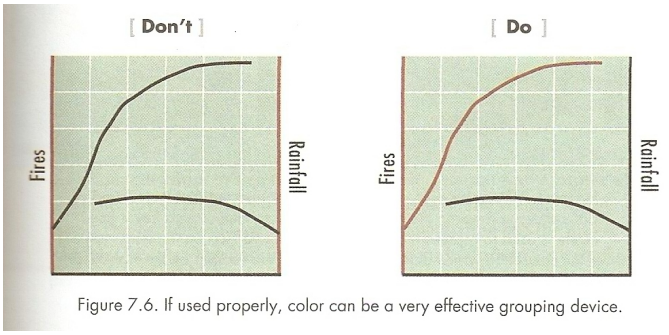


Figure 71: Kosslyn (2006), p. 165, Figure 7.6.

- Avoid using blue if the display is to be photocopied.
- Avoid using hue to represent quantitative information.
- Use deeper saturations and greater intensities for hues that indicate greater amounts.

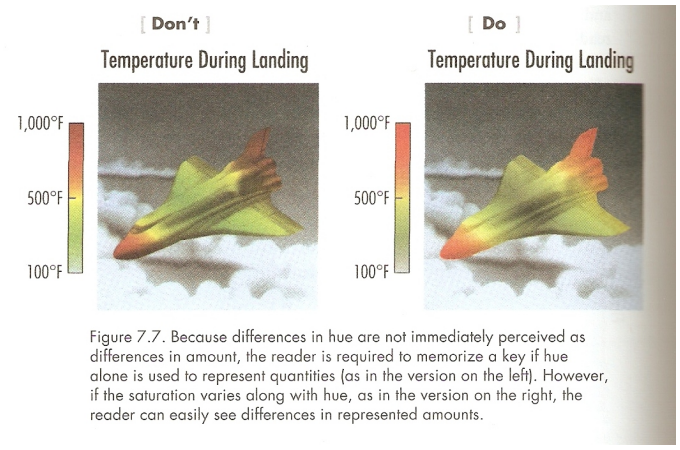


Figure 72: Kosslyn (2006), p. 166, Figure 7.7.

- Do not use hue, saturation, and intensity to specify different measurements.

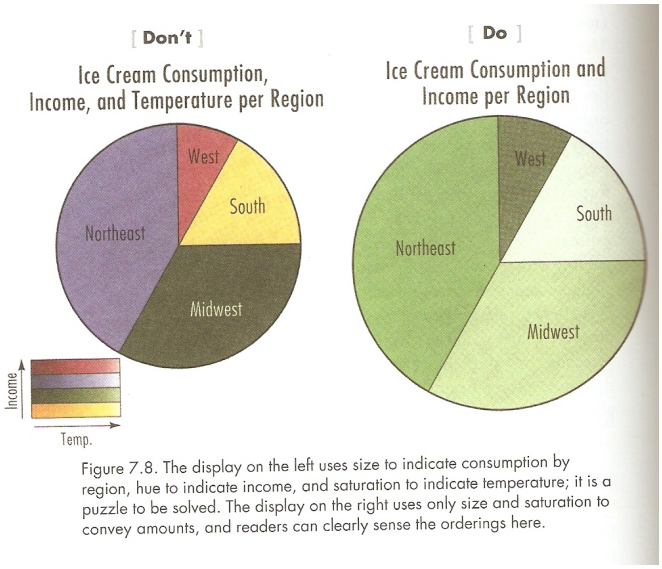


Figure 7.8. The display on the left uses size to indicate consumption by region, hue to indicate income, and saturation to indicate temperature; it is a puzzle to be solved. The display on the right uses only size and saturation to convey amounts, and readers can clearly sense the orderings here.

Figure 73: Kosslyn (2006), p. 168, Figure 7.8.

3.4 Good Color Choices

Compare Figures 74 and 75, taken from Tufte (1997). Both use 21 distinct colors to communicate altitude and ocean depth. Which is better? — Why?

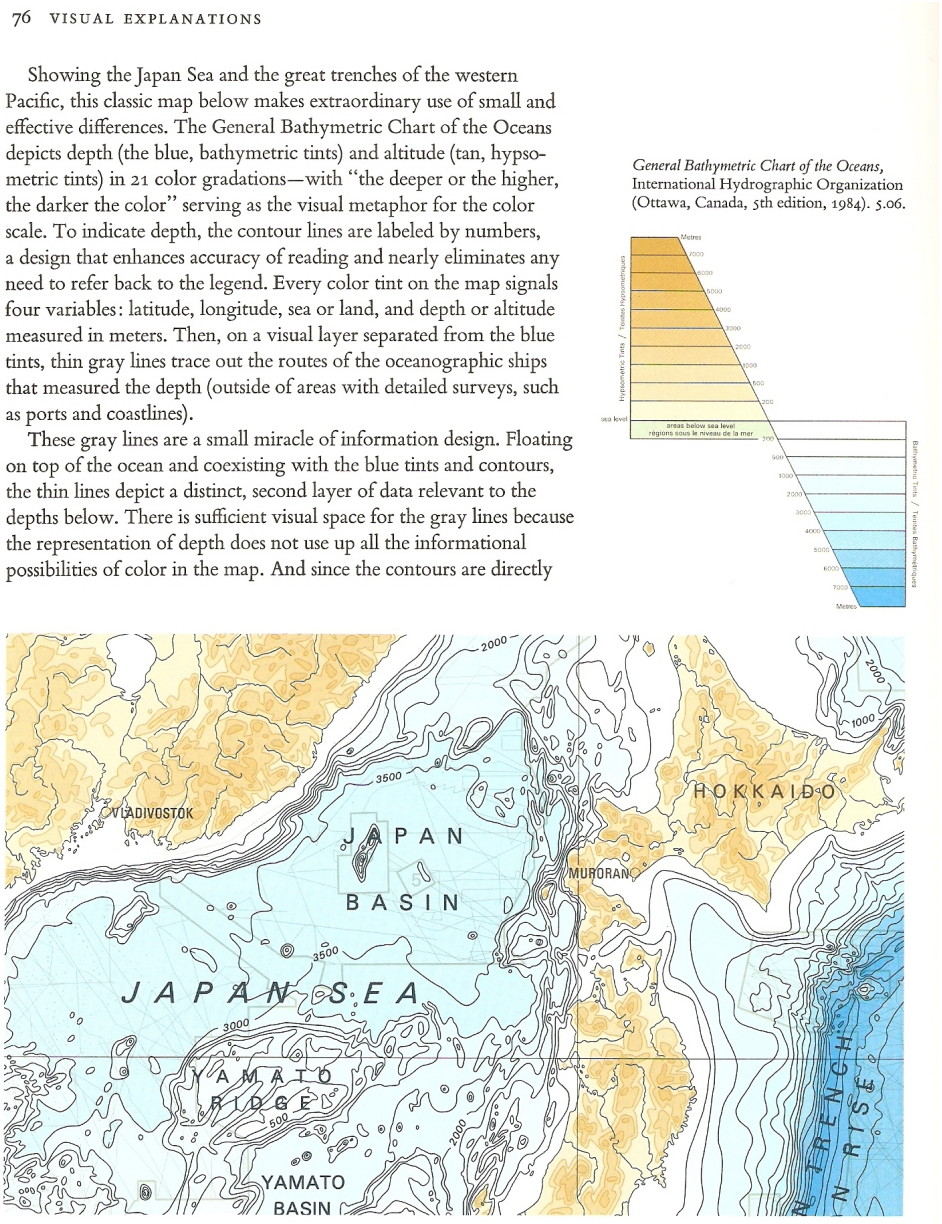


Figure 74: Tufte (1997), p. 76, Figure.

labeled with numbers, the fine distinctions in blue remain clear and readable. By indicating depth with visually minimal gradations in color, the cartographers were able to add an extra two-dimensional layer of gray-line data right on top of the ocean contours. Minimal differences allow more differences.

In ghastly contrast below, a rainbow encodes depth. Although often found in scientific publications, such a visually naive color-scale would be laughed right out of the field (or ocean) of cartography. These aggressive colors, so unnatural and unquantitative, render the map incoherent, with some of the original data now lost in the soup.

Minimal distinctions reduce visual clutter. Small contrasts work to enrich the overall visual signal by increasing the number of distinctions that can be made within a single image; thus design by means of small effective differences helps to increase the resolution of our images. In practice, the appropriate size of small contrasts will depend on the context, priority of particular elements in the overall visual story, number of differentiations made within an image, and characteristics of those viewing the image. Despite these local complications, the global principle of the smallest effective difference resolves many visual issues—serving perhaps even as an algorithm for automated design.

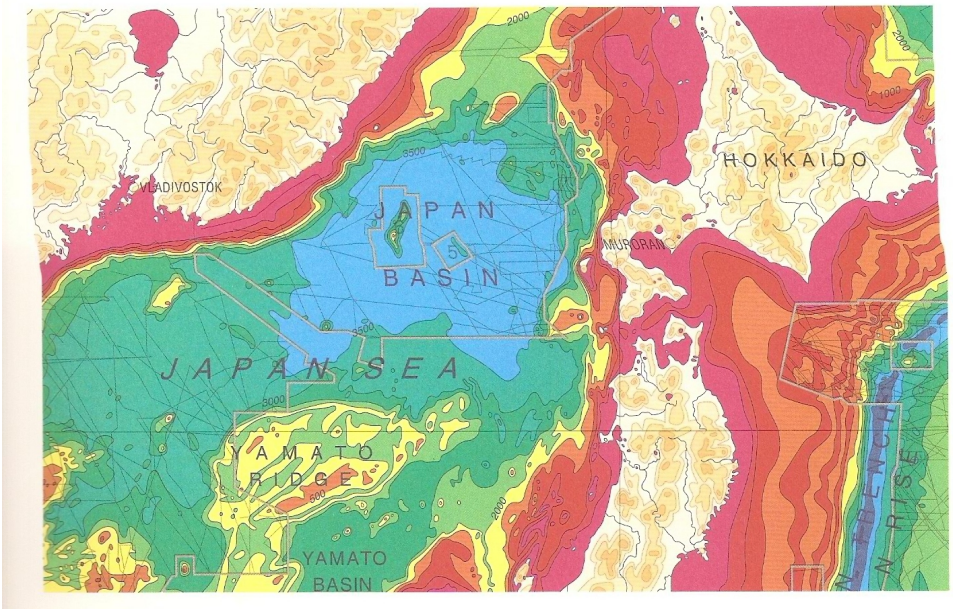


Figure 75: Tufte (1997), p. 77, Figure.

3.4.1 Work by Cindy Brewer and Collaborators

Extensive work regarding the use of color, in particular on maps, was done by Cindy Brewer and her collaborators. Examples of her work include:

- Brewer (1997)
- Brewer et al. (1997)
- Brewer (1999)
- Brewer & Pickle (2002)

A resulting software tool, *ColorBrewer* is described in:

- Leslie (2002), a brief independent review (online version available at `http://www.sciencemag.org/cgi/reprint/296/5567/435c`)
- Harrower & Brewer (2003) (preprint available at `http://www.geography.wisc.edu/~harrower/pdf/ColorBrewer2003.pdf`)
- Brewer (2003)

ColorBrewer is accessible at `http://ColorBrewer.org` and it is described as:

> "ColorBrewer is an online tool designed to help people select good color schemes for maps and other graphics. It is free to use, although we'd appreciate it if you could cite us if you decide to use one of our color schemes."

Main Color Schemes in `http://ColorBrewer.org` are:

sequential: best suited for ordered data that progress from low to high

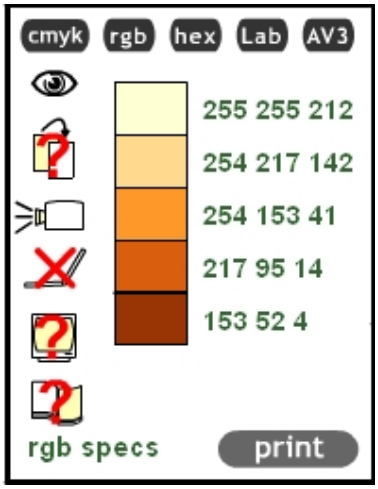


Figure 76: `http://ColorBrewer.org`: 5–class sequential YlOrBr.

diverging: equal emphasis on mid–range critical values and extremes at both ends of the data range

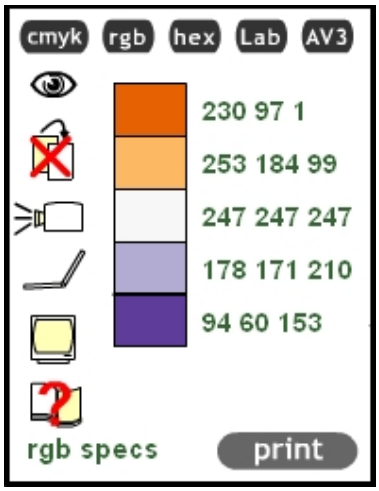


Figure 77: `http://ColorBrewer.org`: 5–class diverging PuOr.

qualitative: no difference in magnitude between legend classes; hues are used to create the primary visual differences; best suited for categorical data

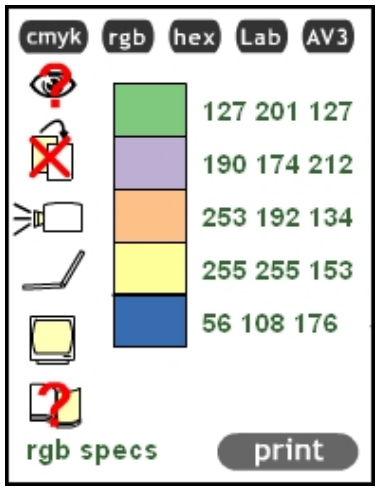


Figure 78: `http://ColorBrewer.org`: 5–class qualitative Accents.

The related R package *RColorBrewer* is documented at `http://cran.r-project.org/web/packages/RColorBrewer/index.html`. First run the examples on pages 3 and 4, then further experiment with these color palettes.

Lecture 17:
Tu 02/17/09

3.4.2 Work by Zeileis, Hornik, and Murrell

In a recent paper, Zeileis et al. (2008) suggest to work with HCL color palettes instead of HSV or RGB palettes.

They use the same distinction among color palettes as in `http://ColorBrewer.org`:

Qualitative Palettes:

Sequential Palettes:

Diverging Palettes:

The related R package *colorspace* is documented at `http://cran.r-project.org/web/packages/colorspace/index.html`. First run the examples on pages 16 and 17, related to *rainbow_hcl*, then further experiment with these color palettes.

3.5 Further Reading

Additional sources for the use of colors in statistical graphics are:

- Tufte (1990), Chapter 5

Lecture 18:
We 02/18/09

4 Categorical Plots

4.1 Which Plot Type to Choose?

Often, there exist many valid options how to display (categorical) data.

Zelazny (2001), p. 12, suggests the following project:

> "Sketch as many charts as you can think of using these data: the more the better."

Percentage of January Sales by Region

	Co. A	Co. B
North	13%	39%
South	35%	6%
East	27%	27%
West	25%	28%

Worksheet

Your Name: _____

Worksheet Answers

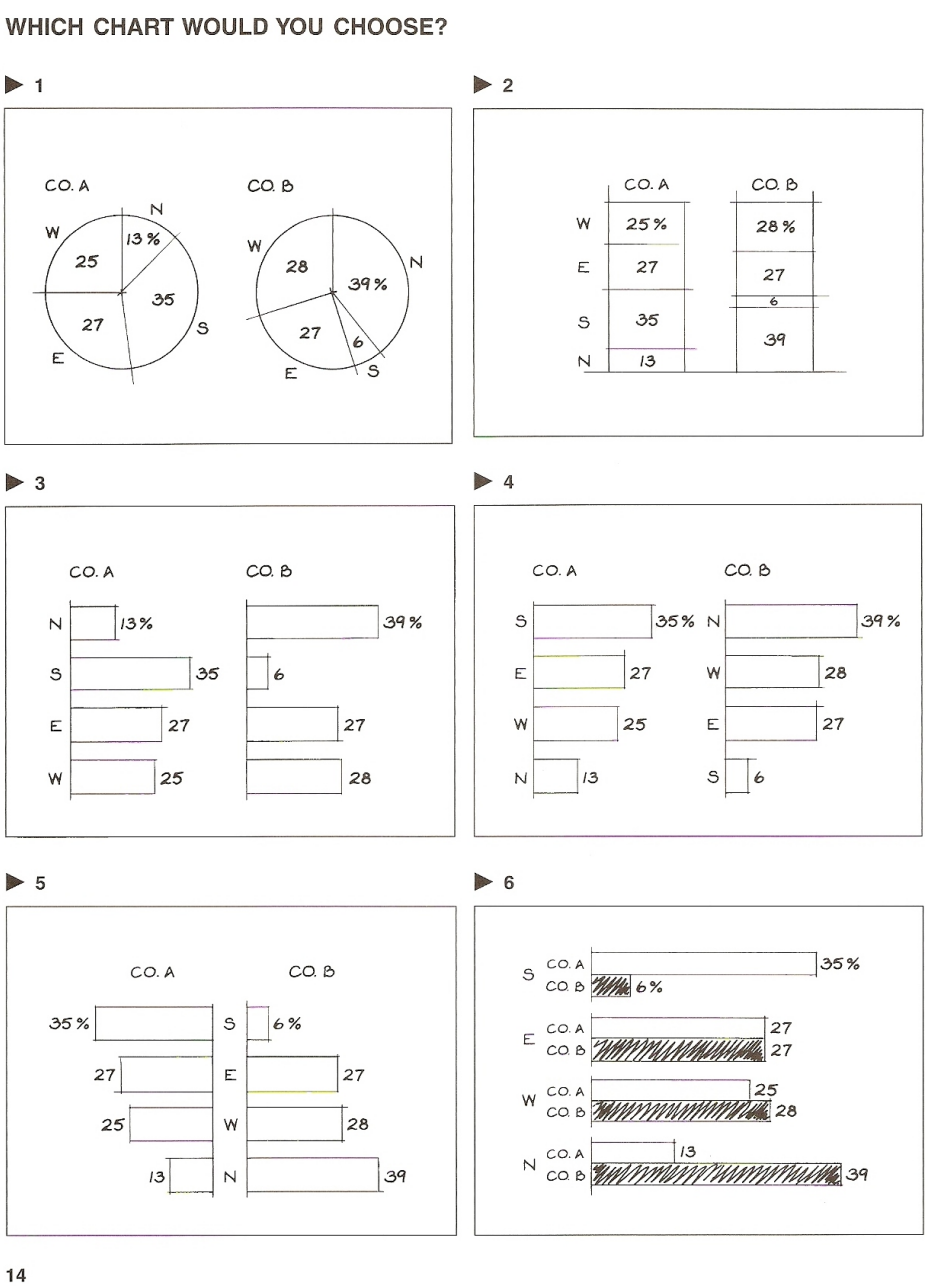


Figure 79: Zelazny (2001), p. 14, Figure.

The charts shown on the facing page may be among those you sketched. All the better if you thought of others. But a question remains.

WHICH CHART WOULD YOU CHOOSE?

It all depends! It all depends on the specific point *you* want to make—*your* message. Each chart shown, simply as a function of the way it's organized, is best equipped to emphasize a particular message.

For instance, showing the data as a couple of pie charts or 100 percent columns, you would be emphasizing that:

▶ **1** ▶ **2** The mix of sales is different for Companies A and B.

Or you may have shown the data as two sets of bar charts, sequencing the bars in the order the data were presented in the table. Now the chart is stressing the message that:

▶ **3** The percentage of sales for both Companies A and B varies by region.

On the other hand, you could have ranked the percentage of sales for each company in descending (or ascending) order, now stressing the point that:

▶ **4** Company A is highest in the South; Company B is highest in the North. Or, Company A is lowest in the North; Company B is lowest in the South.

By structuring the bars in a mirror image around the regions, we now demonstrate that:

▶ **5** Company A's share of sales is highest in the South where Company B's is the weakest.

By grouping the bars against a common base, we now compare the gaps by region, showing that:

▶ **6** In the South, Company A leads B by a wide margin; in the East and West, the two are competitive; in the North, A lags B.

Now, it's possible—even probable—that in the early stages of deciding what your message should be, you may need to sketch a number of charts that look at the data from various points of view. A more efficient approach is to highlight the aspect of the data that seems most important and settle on the message that brings out that aspect.

Figure 80: Zelazny (2001), p. 15, Text.

4.2 Categorical Plots in R

Recall Section 2.4, *"Sex Bias in Graduate Admissions"*, from Freedman et al. (2007), pp. 17–20, many of us are using in our introductory Stat 1040 class.

These data represent aggregate data on applicants to graduate school at Berkeley for the six largest departments in 1973, classified by admission and sex. These data are often used to discuss the issue whether the data show evidence of sex bias in admission practices. There were 2691 male applicants, of whom 1198 (44.5%) were admitted, compared with 1835 female applicants of whom 557 (30.4%) were admitted. Ultimately, this data set is frequently used for illustrating Simpson's paradox and does not show any sex bias when properly analyzed.

In R, the data are stored in a a 3–dimensional array resulting from cross–tabulating 4526 observations on 3 variables. The variables and their levels are as follows:

No	Name	Levels
1	Admit	Admitted, Rejected
2	Gender	Male, Female
3	Dept	A, B, C, D, E, F

In R, this data set is accessible via:

```
UCBAdmissions
```

A better tabular representation can be obtained via:

```
ftable(UCBAdmissions)
```

To obtain the totals as represented in Freedman et al. (2007), p. 18, we have to sum over dimensions 2 and 3 in this 3–dimensional array:

```
apply(UCBAdmissions, c(2, 3), sum)
#
# also, margin.table prodoces the same result
#
margin.table(UCBAdmissions,2:3)
```

To better understand over which dimensions we sum, replace the `c(2, 3)` option with possible other indices, e.g., `1` or `c(1, 2)`. Try a few more.

Question:

How can we calculate in R the percent admitted, as shown in Freedman et al. (2007), p. 18, Table 2? This can be done via a single command line and does not require any loop! And, which single digit do we have to change in our previous R command to obtain the percent rejected?

Table 2. Admissions data for the graduate programs in the six largest majors at University of California, Berkeley.

	Men		*Women*	
Major	*Number of applicants*	*Percent admitted*	*Number of applicants*	*Percent admitted*
A	825	62	108	82
B	560	63	25	68
C	325	37	593	34
D	417	33	375	35
E	191	28	393	24
F	373	6	341	7

Note: University policy does not allow these majors to be identified by name.
Source: The Graduate Division, University of California, Berkeley.

Figure 81: Freedman et al. (2007), p. 18, Table 2.

Answer:

```
# Percent admitted
UCBAdmissions[1, ,] / apply(UCBAdmissions, c(2, 3), sum) * 100
# Percent rejected
UCBAdmissions[2, ,] / apply(UCBAdmissions, c(2, 3), sum) * 100
```

4.2.1 Pie Charts

Let us concentrate on the popularity of the six majors first, i.e., the total number of admissions for each of these majors.

In R, these application numbers can be calculated via:

```
apply(UCBAdmissions, 3, sum)
```

Many people would immediately think of a pie chart as a possible graphical representation:

```
pie(apply(UCBAdmissions, 3, sum))
```

Note that there is no sorting here. Can you easily order the slices by visual inspection, i.e., which major has the largest number/percentage of admissions, which is second, third, etc.?

A better representation is to sort the data from largest to smallest and then plot the slices in clockwise direction, starting with the largest slice at 90∘.

```
pie(sort(apply(UCBAdmissions, 3, sum), decreasing = TRUE),
  clockwise = TRUE,
  main = "UC Berkley Admissions by Major")
```

This is somewhat better, but still not perfect. The R help page for pie charts indicates:

> "Pie charts are a very bad way of displaying information. The eye is good at judging linear measures and bad at judging relative areas. A bar chart or dot chart is a preferable way of displaying this type of data."

Moreover, Cleveland (1985), p. 264, states:

> "Data that can be shown by pie charts always can be shown by a dot chart. This means that judgements of position along a common scale can be made instead of the less accurate angle judgements."

And what about the extremely popular 3D–pie charts that often can be found in business reports and the media? The answer is a clear **Don't.**

Wallgren et al. (1996), p. 70, provide a striking example why not to use 3D–pie charts. Guess the percentages associated with the four different areas:

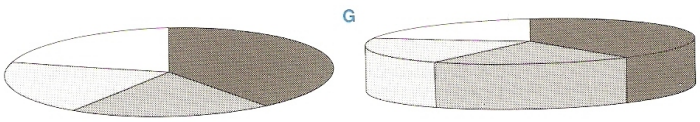


Figure 82: Wallgren et al. (1996), p. 70, Figure G.

And here is the answer:

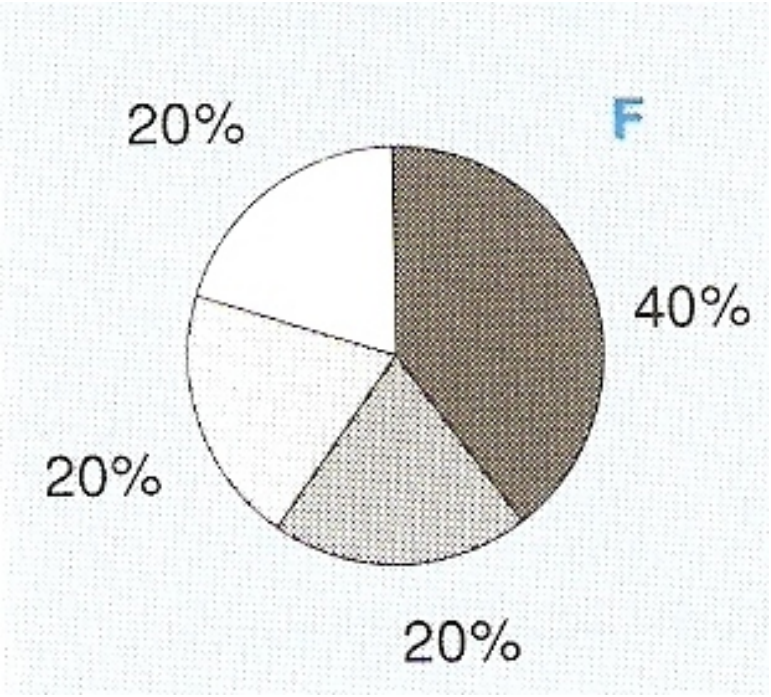


Figure 83: Wallgren et al. (1996), p. 70, Figure F.

4.2.2 Bar Charts

The R help page for barplot indicates:

> "Creates a bar plot with vertical or horizontal bars."

```
UCBAd = margin.table(UCBAdmissions, 1:2)
UCBAd

barplot(UCBAd, legend.text = T)

barplot(UCBAd, legend.text = T, beside = T)
```

The following commands create (divided) bar charts that show the percentage admitted/rejected for each gender.

```
barplot(UCBAd / rbind(margin.table(UCBAd, 2), margin.table(UCBAd, 2)),
  legend.text = T)

barplot(UCBAd / rbind(margin.table(UCBAd, 2), margin.table(UCBAd, 2)),
  legend.text = T, beside = T)
```

Warning:
Cleveland (1994), Section 4.10, "Pop Charts", p. 262, strongly advises against the use of pie charts, divided bar charts, and area charts:

> Three graphical methods — pie charts, divided bar charts, and area charts — are widely used in mass media and business publications but are used far less in science and technology. Because of their use, we will call these graphical methods *pop charts*.
>
> Any data that can be encoded by one of these pop charts can also be encoded by either a dot plot or a multiway dot plot that typically provides far more efficient pattern perception and table look–up than the pop–chart encoding. Interestingly, the better pattern perception results from a detection operation, a phenomenon that has been missed in previous studies of pop charts.

4.2.3 Dot Charts

The R help page for dotchart indicates:

> "Draw a Cleveland dot plot. [...]
> This function is invoked for its side effect, which is to produce two variants of dotplots as described in Cleveland (1985). Dot plots are a reasonable substitute for bar plots."

```
dotchart(UCBAd)

UCBMajor = margin.table(UCBAdmissions, 2:3)
dotchart(UCBMajor)

UCBMajorsort = UCBMajor[, order(UCBMajor[1,], decreasing = TRUE)]
dotchart(UCBMajorsort, color = c("red", "blue"))
```

4.2.4 Mosaic Plots

The R help page for mosaicplot indicates:

> "Plots a mosaic on the current graphics device. [...]
> *shade* a logical indicating whether to produce extended mosaic plots, or a numeric vector of at most 5 distinct positive numbers giving the absolute values of the cut points for the residuals. By default, shade is FALSE, and simple mosaics are created. Using shade = TRUE cuts absolute values at 2 and 4."

```
mosaicplot(UCBAd)

mosaicplot(UCBAd, shade = T)

mosaicplot(UCBAdmissions, shade = T)

mosaicplot(aperm(UCBAdmissions, 3:1), shade = T)
```

Lecture 21:
We 02/25/09

4.2.5 Spine Plots and Spinograms

The R help page for spineplot indicates:

> "Spine plots are a special case of mosaic plots, and can be seen as a generalization of stacked (or highlighted) bar plots. Analogously, spinograms are an extension of histograms."

```
#
# compare the use of this command without () ...
#
spineplot(UCBAd)

#
# ... and with ()
#
(spineplot(UCBAd))

(spineplot(t(UCBAd)))

(spineplot(margin.table(UCBAdmissions, c(3, 2)), main = "Applications at UCB"))

(spineplot(margin.table(UCBAdmissions, c(3, 1)), main = "Admissions at UCB"))
```

4.2.6 Four Fold Plots

The R help page for fourfoldplot indicates:

> "Creates a fourfold display of a 2 by 2 by k contingency table on the current graphics device, allowing for the visual inspection of the association between two dichotomous variables in one or several populations (strata). [...]
> *std* a character string specifying how to standardize the table. Must be one of "margins", "ind.max", or "all.max", and can be abbreviated by the initial letter. If set to "margins", each 2 by 2 table is standardized to equate the

margins specified by margin while preserving the odds ratio. If “ind.max” or “all.max”, the tables are either individually or simultaneously standardized to a maximal cell frequency of 1.”

```
fourfoldplot(UCBAd, std = "a")

fourfoldplot(UCBAd)

fourfoldplot(UCBAdmissions, std = "m")

fourfoldplot(UCBAdmissions, std = "a")
```

4.3 Categorical Plots in Mondrian

According to `http://rosuda.org/mondrian/`:

> "Mondrian is a general purpose statistical data–visualization system. It features outstanding visualization techniques for data of almost any kind, and has its particular strength compared to other tools when working with **Categorical Data, Geographical Data** and **LARGE Data**.
>
> All plots in Mondrian are fully linked, and offer various interactions and queries. Any case selected in a plot in Mondrian is highlighted in all other plots.
>
> Currently implemented plots comprise **Mosaic Plot, Scatterplots and SPLOM, Maps, Barcharts, Histograms, Missing Value Plot, Parallel Coordinates/Boxplots** and **Boxplots y by x**. "

Main references for Mondrian are Theus (2002), Theus (2003), and Theus & Urbanek (2009).

Theus & Urbanek (2009) has an associated Web page at `http://www.interactivegraphics.org`:

> "This site is the web resource for the book "Interactive Graphics for Data Analysis — Principles and Examples".
>
> There are links to the most important software tools, all datasets used in the book for easy download, and a set of slides which may be used together with the book for a lecture.
>
> The R–code used in the book can be found here as well."

4.3.1 Installation

Go to `http://rosuda.org/mondrian/`, then follow the link to the download section on this page. Look over the license condition. If you agree, then download the most recent version of Mondrian (currently 1.0 as of 12/18/2008) by right mouse–clicking on the operating system you use. Save *Mondrian.exe* into a directory of your choice. You can start Mondrian directly (without any additional installation) by mouse–clicking on *Mondrian.exe*.

As a test data set, work with the *Titanic* data available under the *Mondrian Titanic* link or directly from `http://stats.math.uni-augsburg.de/Mondrian/Data/Titanic.txt`. Save these data locally as *Titanic.txt*. Then load them into *Mondrian*.

4.3.2 The Titanic Data in Mondrian

The *Mondrian* description at `http://rosuda.org/mondrian/` indicates:

> "**Titanic**
>
> Data set on the 2201 passengers of the Titanic. Pure categorical with data on class, gender, age and survival."

The interactive exploration of the *Titanic* data via *Mondrian* has been further discussed in Theus & Urbanek (2009), Examples D: The Titanic Disaster Revisited, pp. 183–191.

Task:
Interactively recreate the nine plots from Figure 84 using *Mondrian*.

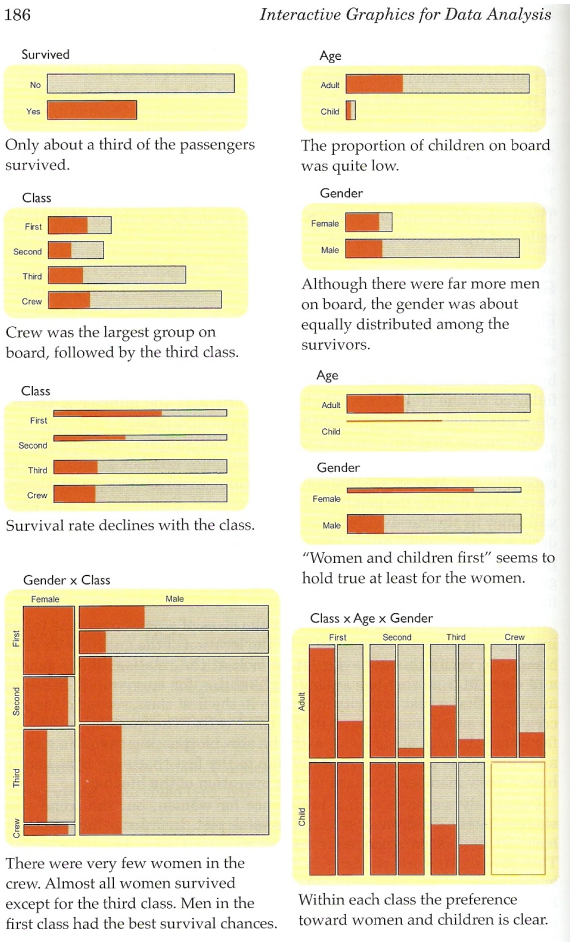


Figure 84: Theus & Urbanek (2009), p. 186, Figure.

Answer:

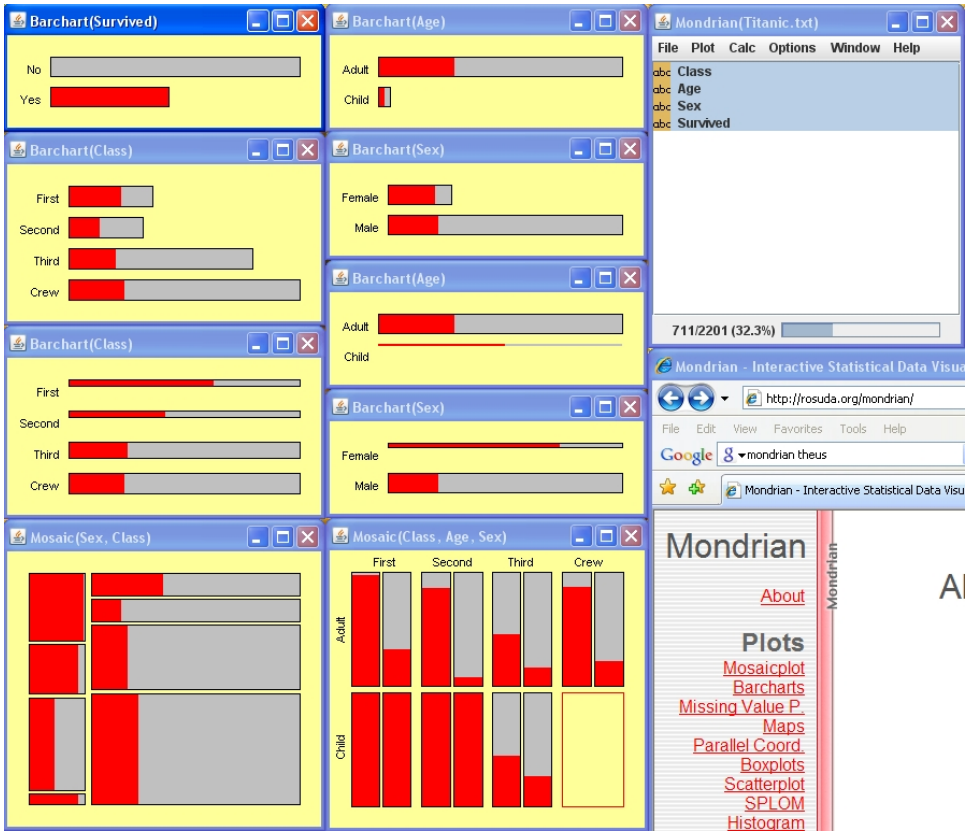


Figure 85: *Mondrian* output related to Theus & Urbanek (2009), p. 186, Figure.

4.4 Further Reading

Additional sources for the visualization of categorical data are:

- Blasius & Greenacre (1998)
- Friendly (2000*b*)
- Hofmann (2007)
- Theus & Urbanek (2009)

4.5 R Code and Output

The automatic output with R code, numerical results, and graphical output in this section and sections in the next chapters are created via the R tool *Sweave*. According to `http://www.statistik.lmu.de/~leisch/Sweave/`:

> "**What is Sweave?**
> Sweave is a tool that allows to embed the R code for complete data analyses in latex documents. The purpose is to create dynamic reports, which can be updated automatically if data or analysis change. Instead of inserting a prefabricated graph or table into the report, the master document contains the R code necessary to obtain it. When run through R, all data analysis output (tables, graphs, etc.) is created on the fly and inserted into a final latex document. The report can be automatically updated if data or analysis change, which allows for truly reproducible research. "

Depending on the hardware platform where you are running *Sweave*, you may have to adjust paths — or you can simply copy the *Sweave.sty* file from the appropriate R directory (e.g., `C:\Program Files\R\R-2.8.1\share\texmf`) into your working directory where the LaTeX file is located.

Then, within R, you can invoke *Sweave* (without having to install anything else) via:

```
Sweave("lect_chapter4_sweave1.snw")
```

When successful, you should obtain a response like this:

```
You can now run LaTeX on 'lect_chapter4_sweave1.tex'
```

A brief overview on *Sweave* can be found at Adele Cutler's Web page at `http://www.math.usu.edu/~adele/s6100/sweave.ppt`. For full details including frequently asked questions, visit the official *Sweave* homepage at `http://www.statistik.lmu.de/~leisch/Sweave/`.

The first version of *Sweave* is described in Leisch (2002). A preprint is available at `http://www.statistik.lmu.de/~leisch/Sweave/Sweave-compstat2002.pdf`.

4.5.1 Example 1: UCBAdmissions

The R description indicates:

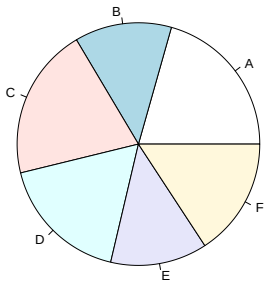
> **Student Admissions at UC Berkeley**
>
> Aggregate data on applicants to graduate school at Berkeley for the six largest departments in 1973 classified by admission and sex.

```
> UCBAdmissions

, , Dept = A

          Gender
Admit      Male Female
  Admitted  512     89
  Rejected  313     19

, , Dept = B

          Gender
Admit      Male Female
  Admitted  353     17
  Rejected  207      8

, , Dept = C

          Gender
Admit      Male Female
  Admitted  120    202
  Rejected  205    391

, , Dept = D

          Gender
Admit      Male Female
  Admitted  138    131
  Rejected  279    244
```

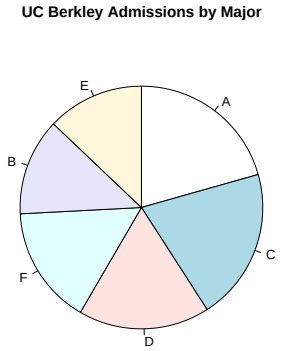
```
, , Dept = E

          Gender
Admit      Male Female
  Admitted   53     94
  Rejected  138    299

, , Dept = F

          Gender
Admit      Male Female
  Admitted   22     24
  Rejected  351    317
```

```
> ftable(UCBAdmissions)

               Dept   A   B   C   D   E   F
Admit    Gender
Admitted Male        512 353 120 138  53  22
         Female       89  17 202 131  94  24
Rejected Male        313 207 205 279 138 351
         Female       19   8 391 244 299 317

> apply(UCBAdmissions, c(2, 3), sum)

        Dept
Gender     A   B   C   D   E   F
  Male   825 560 325 417 191 373
  Female 108  25 593 375 393 341

> #
> # also, margin.table prodoces the same result
> #
> margin.table(UCBAdmissions, 2:3)

        Dept
Gender     A   B   C   D   E   F
  Male   825 560 325 417 191 373
  Female 108  25 593 375 393 341
```

```
> # Percent admitted
> UCBAdmissions[1■] / apply(UCBAdmissions, c(2, 3), sum) * 100

        Dept
Gender          A        B        C        D        E        F
  Male   62.0606 63.03571 36.92308 33.09353 27.74869 5.898123
  Female 82.4074 68.00000 34.06408 34.93333 23.91858 7.038123

> # Percent rejected
> UCBAdmissions[2■] / apply(UCBAdmissions, c(2, 3), sum) * 100

        Dept
Gender          A        B        C        D        E        F
  Male   37.93939 36.96429 63.07692 66.90647 72.25131 94.10188
  Female 17.59259 32.00000 65.93592 65.06667 76.08142 92.96188

> apply(UCBAdmissions, 3, sum)

  A   B   C   D   E   F
933 585 918 792 584 714

> pie(apply(UCBAdmissions, 3, sum))
```

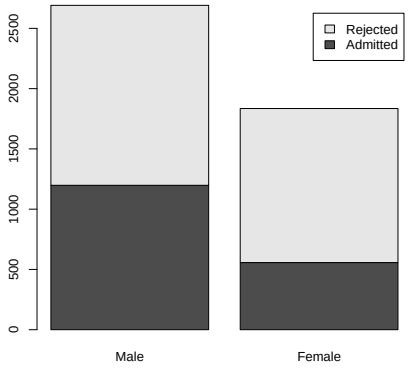



```
> pie(sort(apply(UCBAdmissions, 3, sum), decreasing = TRUE),
+   clockwise = TRUE,
+   main = "UC Berkley Admissions by Major")
```

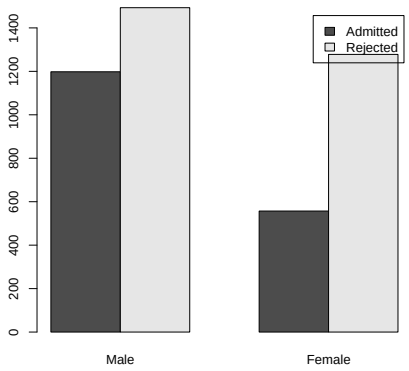



```
> UCBAd = margin.table(UCBAdmissions, 1:2)
> UCBAd

          Gender
Admit      Male Female
  Admitted 1198    557
  Rejected 1493   1278
```

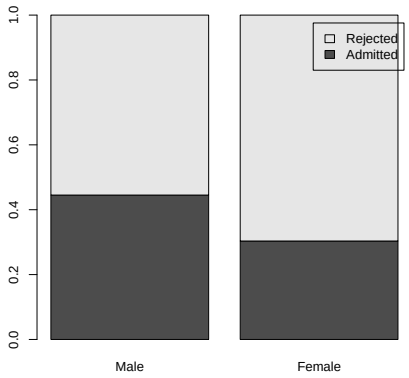
```
> barplot(UCBAd, legend.text = T)
```



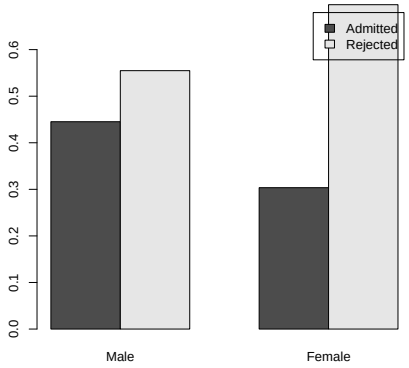

```
> barplot(UCBAd, legend.text = T, beside = T)
```



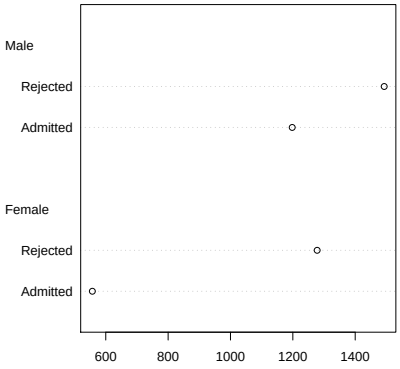

```
> barplot(UCBAd / rbind(margin.table(UCBAd, 2), margin.table(UCBAd, 2)),
+   legend.text = T)
```



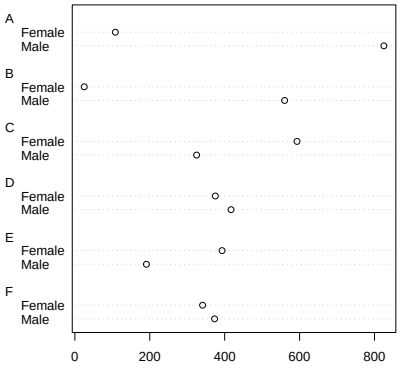

```
> barplot(UCBAd / rbind(margin.table(UCBAd, 2), margin.table(UCBAd, 2)),
+   legend.text = T, beside = T)
```



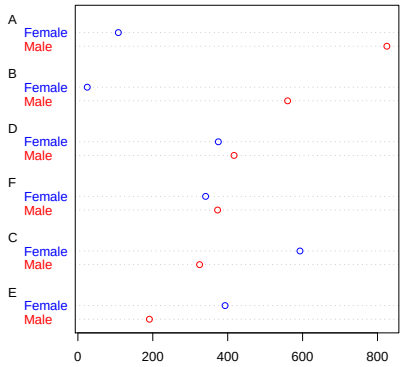

```
> dotchart(UCBAd)
```



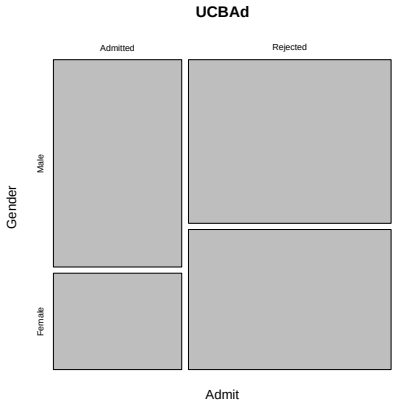

```
> UCBMajor = margin.table(UCBAdmissions, 2:3)
> dotchart(UCBMajor)
```



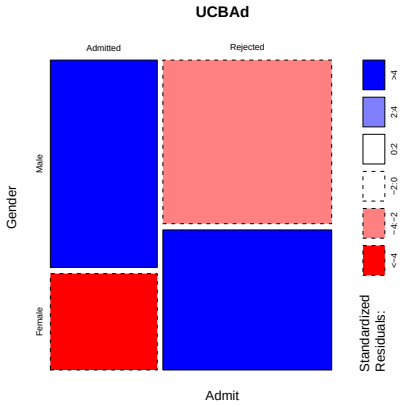
```
> UCBMajorsort = UCBMajor[, order(UCBMajor[1,], decreasing = TRUE)]
> dotchart(UCBMajorsort, color = c("red", "blue"))
```



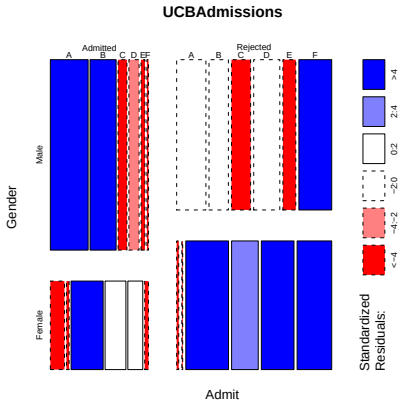

```
> mosaicplot(UCBAd)
```



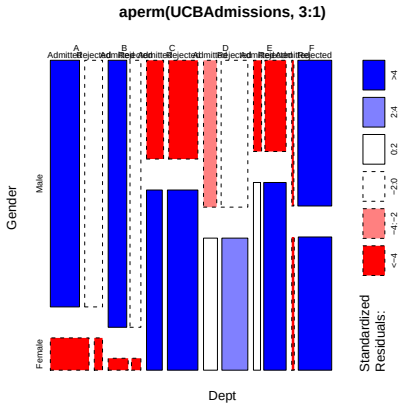

```
> mosaicplot(UCBAd, shade = T)
```



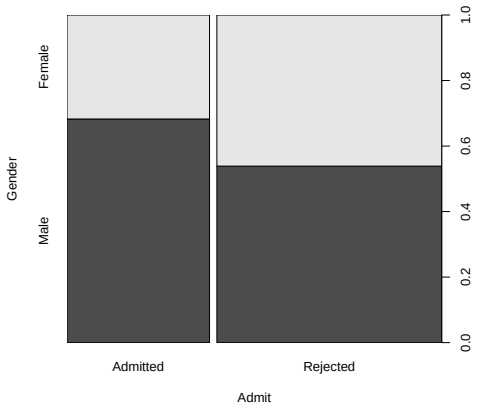

```
> mosaicplot(UCBAdmissions, shade = T)
```




```
> mosaicplot(aperm(UCBAdmissions, 3:1), shade = T)
```

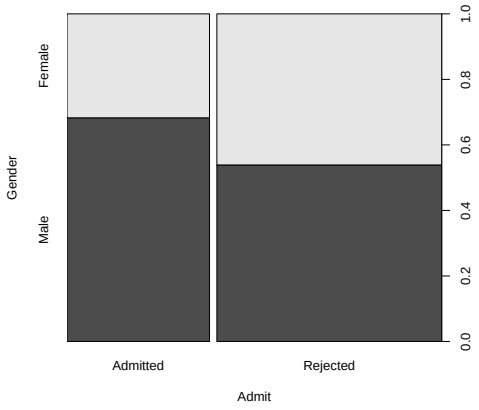


```
> #
> # compare the use of this command without () ...
> #
> spineplot(UCBAd)
```



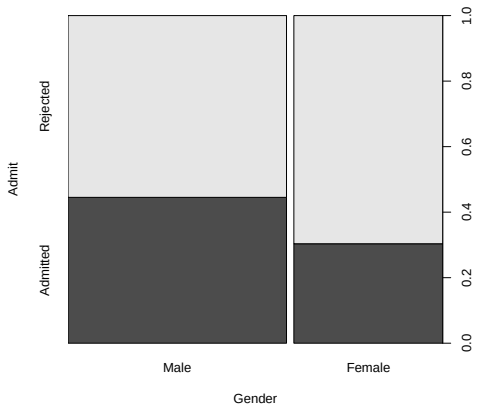

```
> #
> # ... and with ()
> #
> (spineplot(UCBAd))
```

```
          Gender
Admit      Male Female
  Admitted 1198    557
  Rejected 1493   1278
```



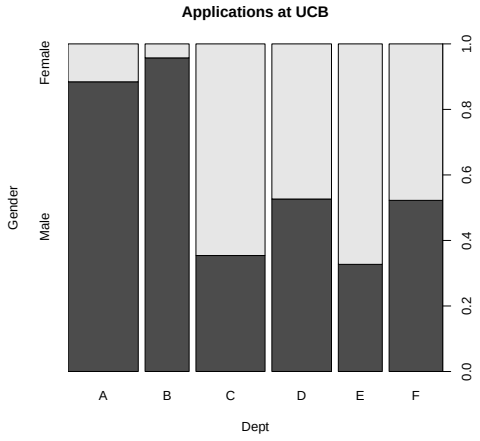
```
> (spineplot(t(UCBAd)))
```

```
        Admit
Gender   Admitted Rejected
  Male       1198     1493
  Female      557     1278
```



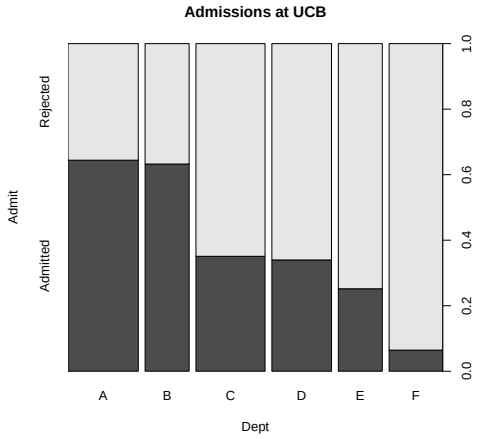

```
> (spineplot(margin.table(UCBAdmissions, c(3, 2)), main = "Applications at UCB"))
```

```
    Gender
Dept Male Female
   A  825    108
   B  560     25
   C  325    593
   D  417    375
   E  191    393
   F  373    341
```

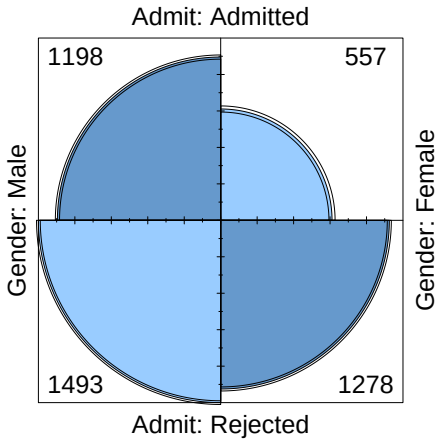


```
> (spineplot(margin.table(UCBAdmissions, c(3, 1)), main = "Admissions at UCB"))
```

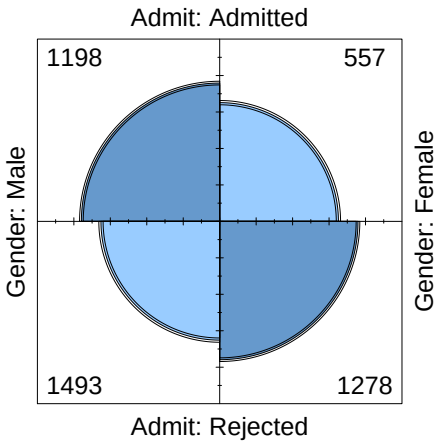
```
    Admit
Dept Admitted Rejected
   A      601      332
   B      370      215
   C      322      596
   D      269      523
   E      147      437
   F       46      668
```



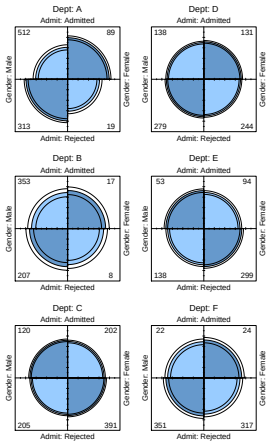

```
> fourfoldplot(UCBAd, std = "a")
```



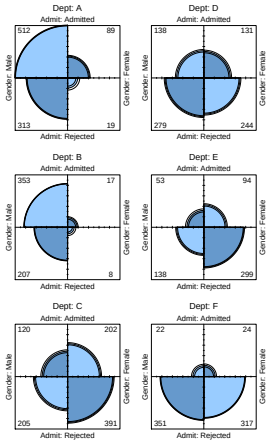

```
> fourfoldplot(UCBAd)
```




```
> fourfoldplot(UCBAdmissions, std = "m")
```




```
> fourfoldplot(UCBAdmissions, std = "a")
```



4.5.2 Example 2: Titanic

The R description indicates:

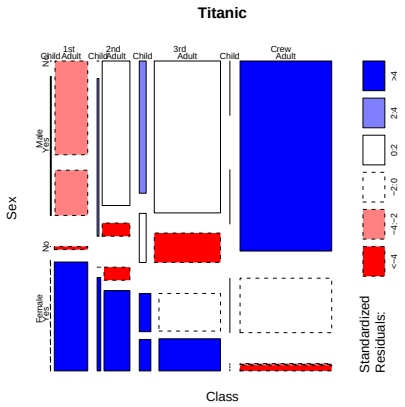
> **Survival of passengers on the Titanic**
>
> This data set provides information on the fate of passengers on the fatal maiden voyage of the ocean liner "Titanic", summarized according to economic status (class), sex, age and survival.

```
> Titanic

, , Age = Child, Survived = No

      Sex
Class  Male Female
  1st     0      0
  2nd     0      0
  3rd    35     17
  Crew    0      0

, , Age = Adult, Survived = No

      Sex
Class  Male Female
  1st   118      4
  2nd   154     13
  3rd   387     89
  Crew  670      3

, , Age = Child, Survived = Yes

      Sex
Class  Male Female
  1st     5      1
  2nd    11     13
  3rd    13     14
  Crew    0      0
```

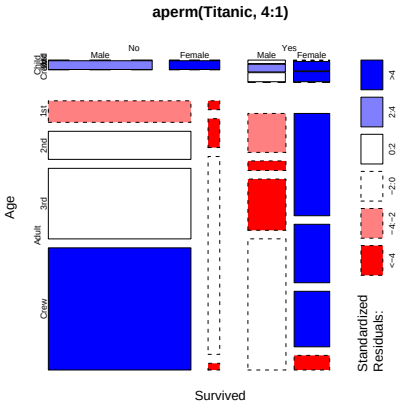
```
, , Age = Adult, Survived = Yes

      Sex
Class  Male Female
  1st    57    140
  2nd    14     80
  3rd    75     76
  Crew  192     20
```

```
> margin.table(Titanic,1)

Class
 1st  2nd  3rd Crew
 325  285  706  885
```

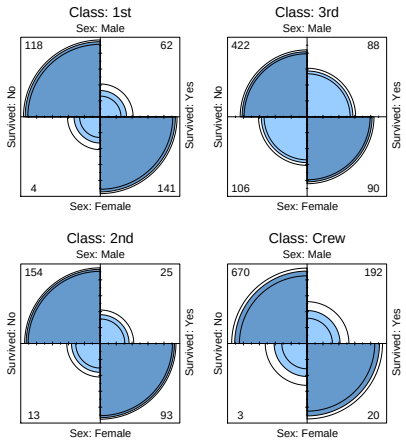
```
> mosaicplot(Titanic, shade = T)
```



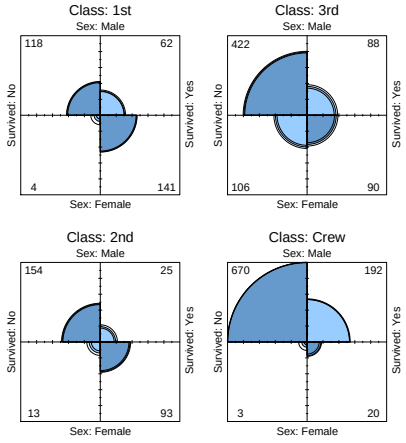

```
> mosaicplot(aperm(Titanic, 4:1), shade = T)
```




```
> Titanic2 <- margin.table(aperm(Titanic, c(2,4,1,3)), 1:3)
> fourfoldplot(Titanic2)
```




```
> fourfoldplot(Titanic2, std = "a")
```

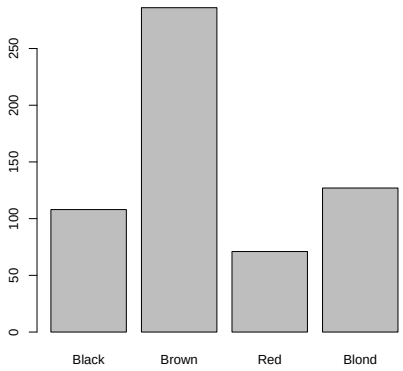


4.5.3 Example 3: HairEyeColor

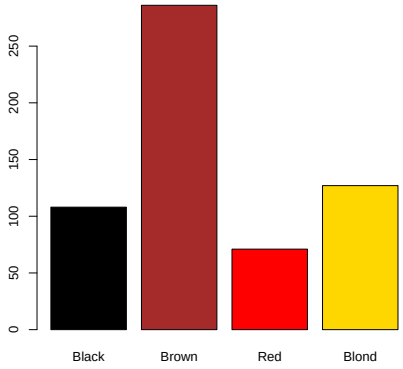
The R description indicates:

> **Hair and Eye Color of Statistics Students**
>
> Distribution of hair and eye color and sex in 592 statistics students.

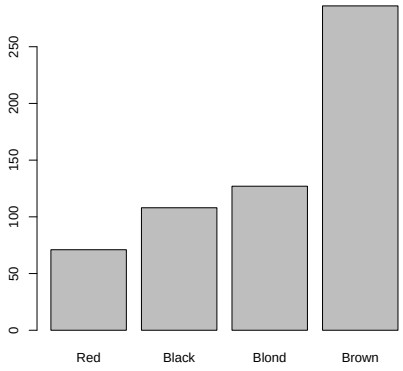
```
> HairEyeColor

, , Sex = Male

       Eye
Hair    Brown Blue Hazel Green
  Black    32   11    10     3
  Brown    53   50    25    15
  Red      10   10     7     7
  Blond     3   30     5     8

, , Sex = Female

       Eye
Hair    Brown Blue Hazel Green
  Black    36    9     5     2
  Brown    66   34    29    14
  Red      16    7     7     7
  Blond     4   64     5     8

> HairCol <- margin.table(HairEyeColor, 1)
> sort(HairCol)

Hair
  Red Black Blond Brown
   71   108   127   286

> barplot(HairCol)
```



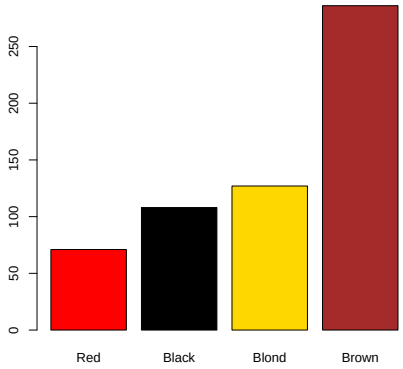

```
> barplot(HairCol, col = c("black", "brown", "red", "gold"))
```



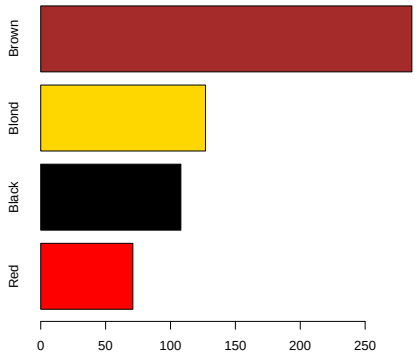

```
> barplot(sort(HairCol))
```



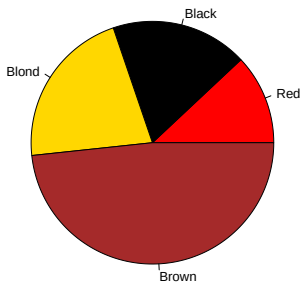
```
> barplot(sort(HairCol), col = c("red", "black", "gold", "brown"))
```




```
> barplot(sort(HairCol), col = c("red", "black", "gold", "brown"), horiz = T)
```



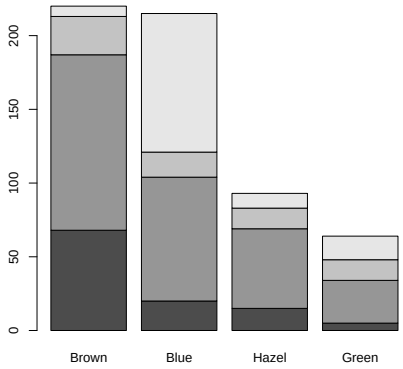

```
> pie(sort(HairCol), col = c("red", "black", "gold", "brown"))
```



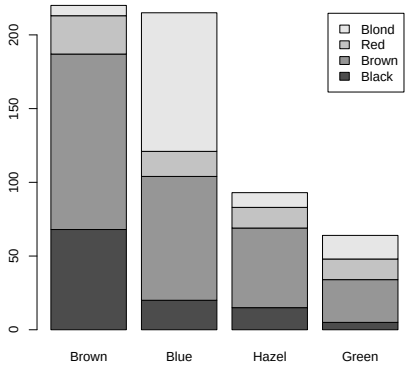
```
> HairEye <- margin.table(HairEyeColor, 1:2)
> HairEye
```

```
       Eye
Hair    Brown Blue Hazel Green
  Black    68   20    15     5
  Brown   119   84    54    29
  Red      26   17    14    14
  Blond     7   94    10    16
```

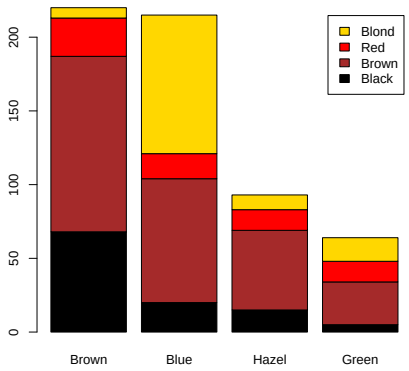
```
> barplot(HairEye)
```



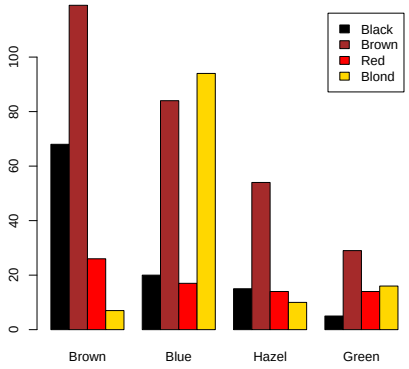

```
> barplot(HairEye, legend.text = T)
```



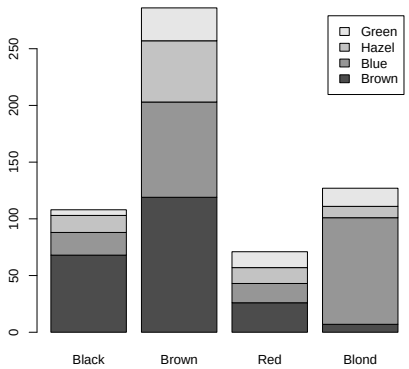

```
> barplot(HairEye, legend.text = T, col = c("black", "brown", "red", "gold"))
```



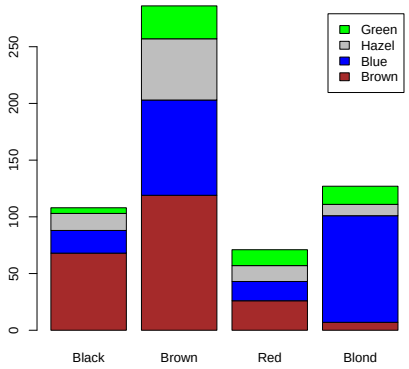

```
> barplot(HairEye, legend.text = T, col = c("black", "brown", "red", "gold"), beside = T)
```



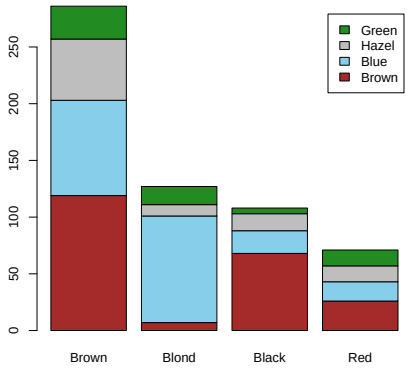

```
> barplot(t(HairEye), legend.text = T)
```



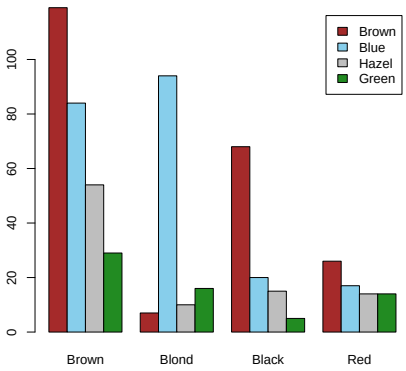
```
> barplot(t(HairEye), legend.text = T, col = c("brown", "blue", "grey", "green"))
```



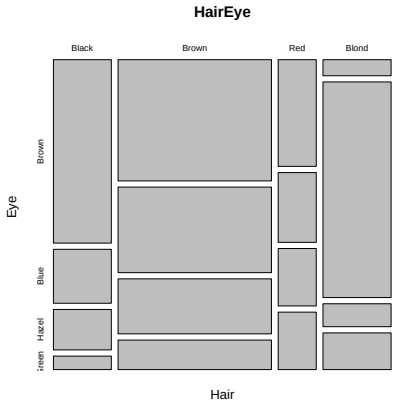

```
> sort(HairCol)

Hair
  Red Black Blond Brown
   71   108   127   286

> barplot(t(HairEye[c(2,4,1,3),]), legend.text = T,
+   col = c("brown", "skyblue", "grey", "forestgreen"))
```



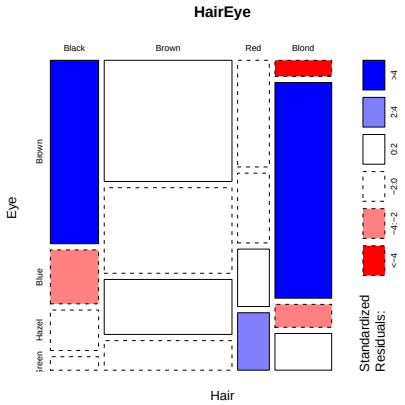

```
> barplot(t(HairEye[c(2,4,1,3),]), legend.text = T,
+   col=c("brown", "skyblue", "grey", "forestgreen"), beside=T)
```



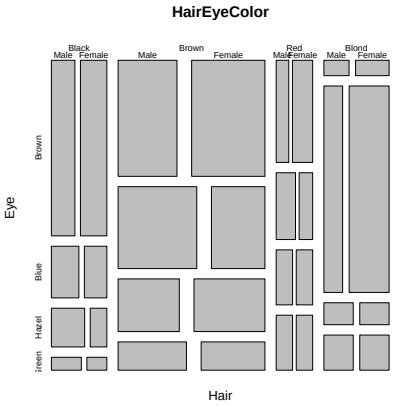

```
> mosaicplot(HairEye)
```



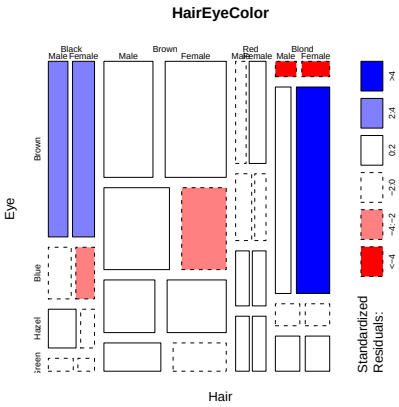

```
> mosaicplot(HairEye, shade = T)
```



```
> mosaicplot(HairEyeColor)
```




```
> mosaicplot(HairEyeColor, shade = T)
```



Lecture 23:
Mo 03/02/09

5 Univariate Plots

5.1 Histograms

Example 1:
Four histograms of the same data set, showing the weights in pounds of 132 professional male athletes.

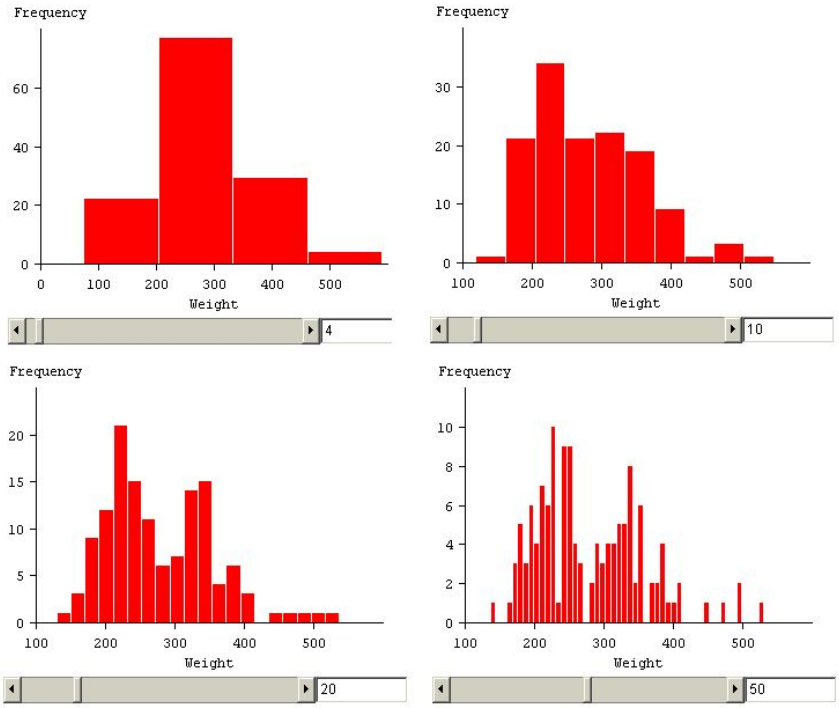


Figure 86: Symanzik, Stat 1040 Lecture Notes, Chapter 3: Four Histograms of the same data set.

Question:
What can we conclude about the underlying data? And which of these four histograms best reveals this fact?

Example 2:

An interactive applet that allows to change the number of classes in a histogram via a slider can be found at `http://www.stat.sc.edu/~west/javahtml/Histogram.html`:

> "**Histogram Applet:**
>
> This applet is designed to teach students how bin widths (or the number of bins) affect a histogram. The histogram below is for the Old Faithful data set. The observations are the duration (in minutes) for eruptions of the Old Faithful geyser in Yellowstone National Park. Students should interactively change the bin width by dragging the arrow underneath the bin width scale. For large bin widths, the bimodal nature of the dataset is hidden, and for small bin widths the plot reduces to a spike at each data point. What bin width do you think provides the best picture of the underlying data?"

The R help page for hist indicates:

> "The generic function hist computes a histogram of the given data values."

The R help page for the Iris data set indicates:

> "This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are *Iris setosa*, *versicolor*, and *virginica*.
>
> iris is a data frame with 150 cases (rows) and 5 variables (columns) named Sepal.Length, Sepal.Width, Petal.Length, Petal.Width, and Species."

Choosing the number of classes for a histogram:

As seen in the previous two examples, a bad choice for the number of classes (nclass or breaks in the R command) in a histogram can almost entirely hide the most interesting information of the underlying data.

Several suggestions for the number of classes exist and are summarized in Venables & Ripley (2002), p. 112. We define range $= x_{(n)} - x_{(1)}$.

- Sturges' formula (default in R):

$$\text{nclass} = \lceil \log_2 n + 1 \rceil, \quad \text{bin width} = \frac{\text{range}}{\text{nclass}},$$

  where $\lceil \ldots \rceil$ indicates the ceiling function.

- Scott's 1979 formula ("scott" in R):

$$\text{bin width} = 3.5\ \hat{\sigma}\ n^{-1/3}, \quad \text{nclass} = \frac{\text{range}}{\text{bin width}},$$

  where $\hat{\sigma}$ is the estimated standard deviation.

- Freedman and Diaconis 1981 formula ("fd" in R):

$$\text{bin width} = 2\ \text{IQR}\ n^{-1/3}, \quad \text{nclass} = \frac{\text{range}}{\text{bin width}},$$

  where IQR is the inter–quartile range.

- Sometimes, the use of nclass $\approx \sqrt{n}$ is suggested:

  `http://www.qimacros.com/qiwizard/how-to-determine-histogram-bin-interval.htm` suggests: *"Take the square root of the number of data points and round up to determine the number of bins required."*

  `http://www.moresteam.com/toolbox/t417.cfm` suggests: *"Calculate the square root of the number of data points and round to the nearest whole number. In the case of our height example, the square root of 50 is 7.07, or 7 when rounded.*

  `http://www.micquality.com/introductory_statistics/int08.htm` states: *"There are various ways of calculating the number of bins. I find that using the square root of the number of data values gives as good a result as the more complicated methods. The value is usually on the low side, but you can adjust it upwards to get convenient bin boundaries. Treat the calculated number of bins as a starting point, and adjust it as necessary to give the result you prefer."*

Example:

```
data(iris)
head(iris)
plength <- iris[,3]
n <- length(plength)

par(mfrow = c(3, 2))

hist(plength, freq = F,
  main = "Default (Sturges) Breaks")
hist(plength, breaks = as.integer(sqrt(n)), freq = F,
  main = "sqrt(n) Breaks")
hist(plength, breaks = "scott", freq = F,
  main = "Scott Breaks")
hist(plength, breaks = "fd", freq = F,
  main = "FD Breaks")

nclass.Sturges(plength)
sqrt(n)
nclass.scott(plength)
nclass.FD(plength)
```

```
h <-3.5 * sd(plength) * n^(-1/3)
h  # bin width
range(plength)
tk <- seq(.9, 8, by = h)
tk
length(tk) - 1  # nclass
hist(plength, breaks = tk, freq = F,
  main = "Exact Scott Breaks")

tk2 <- seq(.9, 8, by = h/2)
tk2
length(tk2) - 1  # nclass
hist(plength, breaks = tk2, freq = F,
  main = "Adjusted Exact Scott Breaks")
```

Finally, how do the histograms for the three species look like?

```
par(mfrow = c(2, 2))

hist(plength, main = "all")

hist(plength[1:50], main = "setosa")

hist(plength[51:100], main = "versicolor")

hist(plength[101:150], main = "virginica")
```

Note:

- These various methods (Sturges, Scott, FD, sqrt) provide suggestions for the number of classes only. To enforce particular breaks, we have to provide a vector giving the exact break points between the histogram cells. However, good software will use the suggestions and then make further adjustments to obtain meaningful class breaks for a human reader, e.g., use integers (and multiples of 5 or 10, etc.) as the boundaries.

- Carefully check whether class intervals are left–open or right–open. R class intervals by default are left–open whereas most readers prefer right–open intervals. Also, check in which class interval the minimum and maximum of a data set are included. For continuous data, there will be little differences in the appearance of a histogram, but for discrete data, different settings may result in a dramatically different visual appearance of a histogram. R provides arguments (include.lowest and right) to adjust these options.

- Wallgren et al. (1996), p. 30, state: **"Since it is relatively complicated both to draw and to read histograms with classes of different size we recommend that, as far as possible, both tables and charts should be made with classes of equal length."**

5.2 Stem–and–Leaf Plots

The R help page for stem indicates:

> "stem produces a stem-and-leaf plot of the values in x."

Venables & Ripley (2002), p. 113, further specify: *"A stem–and–leaf plot is an enhanced histogram. The data are divided into bins, but the 'height' is replaced by the next digits in order."*

```
sort(plength)
```

```
stem(plength)
```

5.3 Boxplots (or Box–and–Whisker Plots)

The R help page for boxplot indicates:

> "Produce box-and-whisker plot(s) of the given (grouped) values.
>
> range: this determines how far the plot whiskers extend out from the box. If range is positive, the whiskers extend to the most extreme data point which is no more than range times the interquartile range from the box. A value of zero causes the whiskers to extend to the data extremes."

The default for range is 1.5.

Venables & Ripley (2002), p. 115, further specify: *"A boxplot is a way to look at the overall shape of a set of data. The central box shows the data between the 'hinges' (roughly quartiles), with the median represented by a line. 'Whiskers' go out to the extremes of the data, and very extreme points are shown by themselves."*

Lecture 24:
We 03/04/09

```
par(mfrow = c(1, 1))

boxplot(plength)

boxplot(plength, range = 0)

boxplot(plength, range = 0.1)

boxplot(plength ~ iris$Species)

boxplot(plength ~ iris$Species, range = 0)

boxplot(plength ~ iris$Species, range = 0.5)
```

5.4 Dot Charts for Univariate Data

The R help page for dotchart indicates:

> "Draw a Cleveland dot plot."

The R help page for UScereal(MASS) indicates:

> "**Nutritional and Marketing Information on US Cereals:**
>
> The UScereal data frame has 65 rows and 11 columns. The data come from the 1993 ASA Statistical Graphics Exposition, and are taken from the mandatory F&DA food label. The data have been normalized here to a portion of one American cup. "

```
library(MASS)  # for cereal data
data(UScereal)
head(UScereal)

Kel.carbs <- UScereal[UScereal$mfr == "K", 7]
Kel.carbs
names(Kel.carbs) <- row.names(UScereal[UScereal$mfr == "K",])
Kel.carbs

dotchart(Kel.carbs)
dotchart(sort(Kel.carbs))
dotchart(sort(Kel.carbs), xlim = c(10, 35),
  xlab = "g carbohydrates per 1 cup serving")
```

Now, activate the lattice package and produce similar graphics:

```
library(lattice)

dotplot(Kel.carbs)  # from lattice library
dotplot(sort(Kel.carbs), xlim = c(10, 35),
  xlab = "g carbohydrates per 1 cup serving")
```

The R help page for barley(lattice) indicates:

> "**Yield data from a Minnesota barley trial:**
>
> Total yield in bushels per acre for 10 varieties at 6 sites in each of two years."

```
library(lattice)  # for barley data
data(barley)
head(barley)

dotplot(variety ~ yield | site, data = barley, groups = year,
  key = simpleKey(levels(barley$year), space = "right"),
  xlab = "Barley Yield (bushels/acre)",
  aspect = 0.5, layout = c(1, 6), ylab = NULL)

levels(barley$site)

# alphabetical sorting of sites (top to bottom)
dotplot(variety ~ yield | site, data = barley, groups = year,
  key = simpleKey(levels(barley$year), space = "right"),
  xlab = "Barley Yield (bushels/acre)",
  aspect = 0.5, layout = c(1, 6), ylab = NULL,
  index.cond = list(c(6,3,4,1,2,5)))
```

Question:
What is the most striking (unusual) feature in these plots? Look carefully!

5.5 Kernel Density Plots for Univariate Data (with Rug Plot)

The R help page for density indicates:

> "**Kernel Density Estimation:**
>
> The (S3) generic function density computes kernel density estimates. Its default method does so with the given kernel and bandwidth for univariate observations. [. . .]
> bw: the smoothing bandwidth to be used. The kernels are scaled such that this is the standard deviation of the smoothing kernel. (Note this differs from the reference books cited below, and from S–PLUS.)
> bw can also be a character string giving a rule to choose the bandwidth. See bw.nrd.
> The specified (or computed) value of bw is multiplied by adjust."

The R help page for rug indicates:

> "Adds a rug representation (1-d plot) of the data to the plot."

Chambers & Hastie (1993), p. 548, further specify: *"rug: [. . .] a univariate histogram or rugplot is displayed along the base of each plot, showing the occurrence of each x–value; ties are broken by jittering."*

```
par(mfrow = c(3, 2))

plot(density(plength), xlim = c(-1, 9))                # default nrd0 bw
rug(plength, ticksize = 0.05)
plot(density(plength, bw = "nrd"), xlim = c(-1, 9))    # normal reference rule bw

plot(density(plength, bw = "ucv"), xlim = c(-1, 9))
                                   # unbiased cross-validation rule bw

plot(density(plength, bw = "bcv"), xlim = c(-1, 9))
                                   # biased cross-validation rule bw

plot(density(plength, bw = "SJ-ste"), xlim = c(-1, 9))
                                   # Sheather-Jones ("solve-the-equation") bw
```

```
plot(density(plength, bw = "SJ-dpi"), xlim = c(-1, 9))
                                    # Sheather-Jones ("direct plug-in") bw

par(mfrow = c(3, 2))

plot(density(plength), xlim = c(-1, 9))
plot(density(plength, adjust = 1/2), xlim = c(-1, 9))  #adjust default bandwidth
plot(density(plength, adjust = 1/4), xlim = c(-1, 9))
plot(density(plength, adjust = 1/8), xlim = c(-1, 9))
plot(density(plength, adjust = 2), xlim = c(-1, 9))
plot(density(plength, adjust = 4), xlim = c(-1, 9))
```

Now, use the lattice package and produce similar graphics:

```
densityplot(plength)  # lattice, takes all parameters from density (above)
densityplot(plength, n = 512)
densityplot(plength, n = 512, bw = "SJ")
```

n is the "number of points at which density is to be evaluated" and the default is 50.

Lecture 26:
Mo 03/16/09

5.6 Quantile–Quantile Plots (Q–Q Plots)

One of the best ways to compare the distribution of a sample $\underline{x}$ of size n with an assumed theoretical distribution F is to use a Quantile–Quantile Plot (Q–Q Plot). In such a plot, we plot the pairs of points

$$\left(F^{-1}\left(\frac{i-0.5}{n}\right), x_{(i)}\right), \quad i = 1, \ldots, n.$$

Example 1:
Convergence of a t distribution with df degrees of freedom towards a normal distribution.

```
# set seed of random number generator to be able to reproduce results
set.seed(1234)

par(mfrow = c(3, 2))

tdf1 = rt(100, df = 1)
qqnorm(tdf1)
qqline(tdf1)

tdf2 = rt(100, df = 2)
qqnorm(tdf2)
qqline(tdf2)

tdf5 = rt(100, df = 5)
qqnorm(tdf5)
qqline(tdf5)

tdf10 = rt(100, df = 10)
qqnorm(tdf10)
qqline(tdf10)

tdf20 = rt(100, df = 20)
qqnorm(tdf20)
```

```
qqline(tdf20)

tdf30 = rt(100, df = 30)
qqnorm(tdf30)
qqline(tdf30)
```

Note:
The closer the points from the sample fall to a straight line, the closer the sample distribution and the theoretical distribution are related. However, here, the greater spread of the extreme quantiles for the sample (for $df = 1, 2, 5, 10, 20$) is an indicator of a long–tailed distribution.

Example 2:
Recall: What is the relationship between exponential distributions and Gamma distributions? Verify this graphically!

```
par(mfrow = c(3, 2))

set.seed(1234)

exp1 = rexp(100, rate = 1)
plot(qgamma(ppoints(exp1), 1, 1), sort(exp1))
abline(0, 1)

exp2 = rexp(100, rate = 2)
plot(qgamma(ppoints(exp2), 1, 2), sort(exp2))
abline(0, 1)

exp5 = rexp(100, rate = 5)
plot(qgamma(ppoints(exp5), 1, 5), sort(exp5))
abline(0, 1)

exp10 = rexp(100, rate = 10)
plot(qgamma(ppoints(exp10), 1, 10), sort(exp10))
abline(0, 1)

# Some major misspecifications
```

```
# a) Swapping shape and rate parameters
plot(qgamma(ppoints(exp2), 2, 1), sort(exp2))
abline(0, 1)

# b) Using 1/rate instead
plot(qgamma(ppoints(exp2), 1, 1/2), sort(exp2))
abline(0, 1)
```

Example 3:

Compare iris plength sample data with an underlying assumed normal distribution. For which of the species is the assumption of normality justified?

```
par(mfrow = c(2, 2))

qqnorm(plength)

qqnorm(plength[1:50])

qqnorm(plength[51:100])

qqnorm(plength[101:150])
```

5.7 Empirical Cumulative Distribution Functions (ECDFs)

Recall from Stat 6720:

Definition 7.1.3:
Let $X_1, \ldots, X_n$ be a sample of size n from a population with distribution F. The function

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty, x]}(X_i)$$

is called **empirical cumulative distribution function** (**empirical cdf, ECDF**). ■

Theorem 7.1.7: **Glivenko–Cantelli Theorem**
$\hat{F}_n(x)$ converges uniformly to $F(x)$, i.e., it holds for all $\epsilon > 0$ that

$$\lim_{n \to \infty} P(\sup_{-\infty < x < \infty} \mid \hat{F}_n(x) - F(x) \mid > \epsilon) = 0.$$

■

Verify this theorem for samples from a normal distribution:

```
par(mfrow = c(2, 2))

set.seed(1234)

xvals = seq(-4, 4, 0.01)

norm10 = rnorm(10)
plot(ecdf(norm10), xlim = c(-4, 4))
lines(xvals, pnorm(xvals))

norm25 = rnorm(25)
plot(ecdf(norm25), xlim = c(-4, 4))
lines(xvals, pnorm(xvals))

norm100 = rnorm(100)
plot(ecdf(norm100), xlim = c(-4, 4))
lines(xvals, pnorm(xvals))
```

```
norm1000 = rnorm(1000)
plot(ecdf(norm1000), xlim = c(-4, 4))
lines(xvals, pnorm(xvals))


# Animation

par(mfrow = c(1, 1))

normgrow = NULL
for (i in 1:20)
{
  normgrow = c(normgrow, rnorm(10))
  plot(ecdf(normgrow), xlim = c(-4, 4), main = length(normgrow))
  lines(xvals, pnorm(xvals))
  Sys.sleep(1)
}
```

Example:
And here the ECDF for iris plength.

```
plot(ecdf(plength))
```

Lecture 27:
We 03/18/09

5.8 Graphics and Small Sample Sizes

Worksheet

Your Name: _____

Question:

The data shown in these four histograms originate from which distribution?

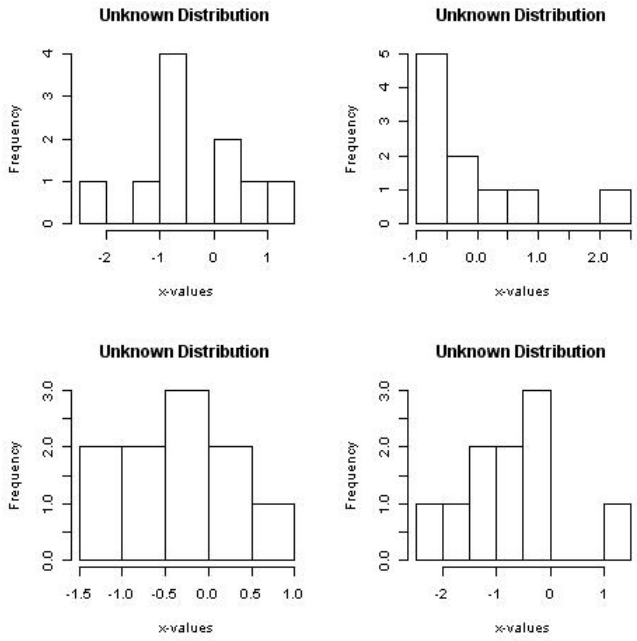


Figure 87: Histograms for data from four unknown distributions.

The corresponding distributions are:

Upper left: _____

Upper right: _____

Lower left: _____

Lower right: _____

Worksheet

Your Name: _____

Question:

Do the data shown in these four qqplots follow a normal distribution?

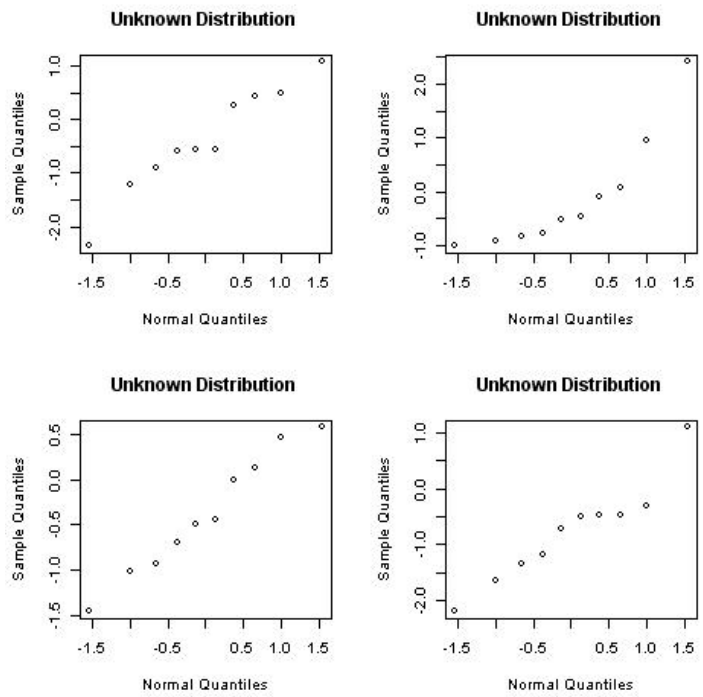


Figure 88: Normal QQplots for data from four unknown distributions.

So, does a particular qqplot suggest that the data originate from a normal distribution? Circle your answer:

Upper left: **yes** / **no**

Upper right: **yes** / **no**

Lower left: **yes** / **no**

Lower right: **yes** / **no**

Answers:

The following R code was used to create the figures shown on the previous two pages:

```
jpeg("Chapter5_unknown_hist.jpg")

par(mfrow = c(2, 2))
set.seed(1234)
xvect1 = NULL

for(i in 1:4)
{
  x = rnorm(10)
  xvect1 = c(xvect1, x)
  hist(x, main = "Unknown Distribution",
    xlab = "x-values")
}

dev.off()


jpeg("Chapter5_unknown_qqplot.jpg")

par(mfrow = c(2, 2))
set.seed(1234)
xvect2 = NULL

for(i in 1:4)
{
  x = rnorm(10)
  xvect2 = c(xvect2, x)
  qqnorm(x, main = "Unknown Distribution",
    xlab = "Normal Quantiles")
}

dev.off()
```

When we jointly plot all 40 observations, we start to see that the underlying distribution indeed is a normal distribution. In fact, all eight plots on the previous two pages show 10 samples each drawn from the standard normal distribution!

```
# Plot all data combined

par(mfrow = c(1, 1))
hist(xvect1)

par(mfrow = c(1, 1))
qqnorm(xvect2)
```

5.9 Further Reading

Additional sources for Trellis Graphics are:

- `http://cm.bell-labs.com/cm/ms/departments/sia/project/trellis/index.html`
- Murrell (2006), Chapter 4
- William G. Jacoby's Web page on Dot Plots: `http://polisci.msu.edu/jacoby/research/dotplots/dotlist.htm`

6 Bivariate Plots

6.1 Scatterplots

Wallgren et al. (1996), p. 46, state:

> "Scatterplots are used to show relationship (causal relationship or covariance) between two *quantitative* variables. The data consists of a number of pairs of coordinates (x, y). Each pair of coordinates is indicated by a dot or circle in a system of coordinates."

In the next example, we resume work with the iris data introduced in the previous chapter.

Example 1:

```
data(iris)
head(iris)
slength <- iris[,1]
swidth <- iris[,2]
species <- iris[,5]

par(mfrow = c(2, 2))

plot(slength, swidth,
  xlim = c(4.0, 8.0), ylim = c(2.0, 5.0))

plot(slength, swidth, pch = 21, bg = unclass(species),
  xlim = c(4.0, 8.0), ylim = c(2.0, 5.0))
legend("topright", levels(species), pch = 21, pt.bg = 1:3)

plot(jitter(slength), jitter(swidth), pch = 22, bg = unclass(species),
  xlim = c(4.0, 8.0), ylim = c(2.0, 5.0))
legend("topright", levels(species), pch = 22, pt.bg = 1:3)

plot(jitter(slength, 5), jitter(swidth, 5), pch = unclass(species),
  xlim = c(4.0, 8.0), ylim = c(2.0, 5.0))
legend("topright", levels(species), pch = 1:3)
```

In the next example, we look at samples of sizes 10, 100, 1,000, 10,000, 100,000, and 1,000,000. from a bivariate normal distribution with correlation $\rho = 0$. Up to 1,000 observations, we can still identify regions of higher and lower density, but for 10,000 (and more) observations, we just have a black center in the plot. We will need some different graphical representation when we have too many observations (and thus too much overplotting) for an ordinary scatterplot.

Example 2:

```
set.seed(1234)

par(mfrow = c(2, 3), pty = "s")

x10 = rnorm(10)
y10 = rnorm(10)
plot(x10, y10,
  xlim = c(-4.0, 4.0), ylim = c(-4.0, 4.0))

x100 = rnorm(100)
y100 = rnorm(100)
plot(x100, y100,
  xlim = c(-4.0, 4.0), ylim = c(-4.0, 4.0))

x1000 = rnorm(1000)
y1000 = rnorm(1000)
plot(x1000, y1000,
  xlim = c(-4.0, 4.0), ylim = c(-4.0, 4.0))

x10000 = rnorm(10000)
y10000 = rnorm(10000)
plot(x10000, y10000,
  xlim = c(-4.0, 4.0), ylim = c(-4.0, 4.0))

x100000 = rnorm(100000)
y100000 = rnorm(100000)
plot(x100000, y100000,
  xlim = c(-4.0, 4.0), ylim = c(-4.0, 4.0))
```

```
x1000000 = rnorm(1000000)
y1000000 = rnorm(1000000)
plot(x1000000, y1000000,
  xlim = c(-4.0, 4.0), ylim = c(-4.0, 4.0))
```

6.2 Hexagon Binning

(Based on Carr et al. (1987))

The R help page for hexbin indicates:

> "Creates a "hexbin" object. Basic components are a cell id and a count of points falling in each occupied cell."

Let us apply hexbin to data from our bivariate normal distribution for 10,000 and 1,000,000 observations.

Example 3:

```
library(hexbin)
library(colorspace)

plot(hexbin(x10000, y10000))

plot(hexbin(x10000, y10000), colramp = heat.colors)

plot(hexbin(x10000, y10000), style = "lattice")

plot(hexbin(x10000, y10000), style = "centroid")

plot(hexbin(x1000000, y1000000), colramp = terrain.colors)

plot(hexbin(x1000000, y1000000, xbins = 16), colramp = topo.colors)
```

Lecture 28:
Fr 03/20/09

6.3 Bivariate Histograms

Execute this code once to create the bivhist3d function, written by Mike Minnotte:

```
bivhist3d <- function(x, y, nbins = nclass.Sturges(x),
  col = c(heat.colors(19)), xlab = "x", ylab = "y")

# Bivariate histogram plot.  Uses the gplots and rgl libraries.
# Written by Mike Minnotte
{
h2d <- hist2d(x, y, nbins, show = F)
xb <- h2d$x
yb <- h2d$y
count <- h2d$counts
xs <- (xb[2]-xb[1])/2
ys <- (yb[2]-yb[1])/2
xb <- (xb+xs)
yb <- (yb+ys)

if (length(col)>1)
  zcol <- matrix(cut(count, length(col), labels=F), length(xb))
else
  zcol <- matrix(1, length(xb), length(yb))

cube <- cube3d()
clear3d()
bg3d(col="grey")

xr <- (max(xb) - min(xb))
yr <- (max(yb) - min(yb))
zr <- (max(count))
mr <- max(xr, yr, zr)*.8

aspect3d(mr/xr, mr/yr, mr/zr)

for (i in 1:length(xb))
  for (j in 1:length(yb))
   {
      scube <- scale3d(cube, xs, ys, count[i,j]/2)
      tcube <- translate3d(scube, xb[i], yb[j], count[i,j]/2)
      if (count[i,j] == 0)
        shade3d(tcube, color = "black")
      else
        shade3d(tcube, color = col[zcol[i,j]])
   }

axes3d()
title3d(xlab = xlab, ylab = ylab, zlab = "count")
return()
}
```

Let us apply bivhist3d to data from our bivariate normal distribution for 10,000 observations.

```
library(gplots)
library(rgl)

bivhist3d(x10000, y10000, col = "red")

bivhist3d(x10000, y10000)

bivhist3d(x10000, y10000, col = terrain.colors(15))
```

If you want to create a snapshot of an interesting 3D view, use the command:

```
rgl.snapshot("bivhist1.png")
```

Finally, let us do some automatic animation:

```
start <- proc.time()[3]
while ((i <- 36*(proc.time()[3] - start)) < 360)
{
  rgl.viewpoint(i, 75, fov = 10)
}
```

Lecture 29:
Mo 03/23/09

7 Trivariate Plots

7.1 Scatterplot Matrix

The R help page for pairs indicates:

> "A matrix of scatterplots is produced. [...] The ijth scatterplot contains $x[, i]$ plotted against $x[, j]$."

The R help page for states indicates:

> "Data sets related to the 50 states of the United States of America. [...]
> state.x77: matrix with 50 rows and 8 columns giving the following statistics in the respective columns.
>
> | | |
> |---|---|
> | Population: | population estimate as of July 1, 1975 |
> | Income: | per capita income (1974) |
> | Illiteracy: | illiteracy (1970, percent of population) |
> | Life Exp: | life expectancy in years (1969–71) |
> | Murder: | murder and non–negligent manslaughter rate per 100,000 population (1976) |
> | HS Grad: | percent high–school graduates (1970) |
> | Frost: | mean number of days with minimum temperature below freezing (1931–1960) in capital or large city |
> | Area: | land area in square miles" |

Example 1:

```
data(state)
head(state.x77)

tristate <- state.x77[, c(3,6,5)] # trivariate data - illiteracy, hs grad, murder

illiteracy <- tristate[,1]
grad <- tristate[,2]
murder <- tristate[,3]
```

```
pairs(tristate)  # pairwise scatterplot
pairs(tristate, col = unclass(state.region))
pairs(tristate, pch = 21, bg = unclass(state.region))
pairs(tristate, pch = 21, bg = unclass(state.region), cex = 2)

pairs(tristate, pch = 21, bg = unclass(state.region), panel = panel.smooth)
```

7.2 3D Scatterplots

Example 2:

```
## lattice

library(lattice)

cloud(murder~illiteracy*grad)


## scatterplot3d

library(scatterplot3d)

scatterplot3d(tristate)

scatterplot3d(tristate, highlight.3d = T)


## rgl

library(rgl)

plot3d(tristate)

plot3d(tristate, type = "s")

plot3d(tristate, type = "s", radius = 0.5)
```

```
plot3d(tristate, type = "h")

plot3d(tristate, type = "s", radius = 0.5, col = "red")

plot3d(tristate, type = "s", radius = 0.5,
  col = c("red", "yellow", "blue", "green")[unclass(state.region)])

text3d(illiteracy, grad, murder + 1.0, text = state.name)
```

Example 3:
The R help page for randu indicates:

> "Random Numbers from Congruential Generator RANDU: 400 triples of successive random numbers were taken from the VAX FORTRAN function RANDU running under VMS 1.5."

```
data(randu)
head(randu)

pairs(randu)

plot3d(randu, type = "s", radius = 0.025, col = "red")
```

7.3 Co–Plots

Cleveland (1993), p. 182, states:

> "The concept of *conditioning* [...] is a fundamental one that forms the basis of a number of graphical methods developed in the past [...]. And it forms the basis for the *conditioning plot*, or *coplot*, a particularly powerful visualization tool for studying how a response depends on two or more factors. "

Example 4:

```
# create dataframe
tristatenew = matrix(c(murder, illiteracy, grad), 50, 3)
colnames(tristatenew) = c("murder", "illiteracy", "grad")
tristatenewdf = as.data.frame(tristatenew)

coplot(murder~illiteracy|grad, data = tristatenewdf)

coplot(grad~illiteracy|murder, data = tristatenewdf)

m.interval = co.intervals(tristatenewdf$murder, number = 6, overlap = 0)
coplot(grad~illiteracy|murder, data = tristatenewdf, given.values = m.interval)

m.interval = co.intervals(tristatenewdf$murder, number = 6, overlap = 0.25)
coplot(grad~illiteracy|murder, data = tristatenewdf, given.values = m.interval)

m.interval = co.intervals(tristatenewdf$murder, number = 4, overlap = 0)
coplot(grad~illiteracy|murder, data = tristatenewdf, given.values = m.interval)

m.interval = co.intervals(tristatenewdf$murder, number = 4, overlap = 0)
coplot(grad~illiteracy|murder, data = tristatenewdf, given.values = m.interval,
  rows = 1)

m.interval = co.intervals(tristatenewdf$murder, number = 9, overlap = 0)
coplot(grad~illiteracy|murder, data = tristatenewdf, given.values = m.interval)

m.interval = co.intervals(tristatenewdf$murder, number = 4, overlap = 0)
coplot(grad~illiteracy|murder, data = tristatenewdf, given.values = m.interval,
  panel = function(x, y, ...) panel.smooth(x, y, span = 0.3, ...))

m.interval = co.intervals(tristatenewdf$murder, number = 4, overlap = 0)
coplot(grad~illiteracy|murder, data = tristatenewdf, given.values = m.interval,
  panel = function(x, y, ...) panel.smooth(x, y, span = 0.7, ...))

m.interval = co.intervals(tristatenewdf$murder, number = 4, overlap = 0)
coplot(grad~illiteracy|murder, data = tristatenewdf, given.values = m.interval,
  panel = function(x, y, ...) panel.smooth(x, y, span = 1.5, ...))
```

Lecture 30:
We 03/25/09

7.4 Trivariate Density Estimation

Execute this code once to create the tkde (three–dimensional kernel density estimation) function, written by Mike Minnotte:

```
tkde <- function(x, y=NULL, z=NULL, hx=NULL, hy=NULL, hz=NULL,
xlist=NULL, ylist=NULL, zlist=NULL, extend=.1,
nx=20, ny=20, nz=20, adjust=c(1,1,1),
xlab=NULL, ylab=NULL, zlab=NULL,...)

{data <- xyz.coords(x,y,z)
x <- data$x
y <- data$y
z <- data$z

n <- length(x)

if (is.null(hx)) hx <- bw.SJ(x,method="dpi")
if (is.null(hy)) hy <- bw.SJ(y,method="dpi")
if (is.null(hz)) hz <- bw.SJ(z,method="dpi")

if (is.null(xlist) )
{dx <- (max(x)-min(x))*extend
xlist <-seq(min(x)-dx,max(x)+dx,length=nx)}
else nx <- length(xlist)

if (is.null(ylist) )
{dy <- (max(y)-min(y))*extend
ylist <-seq(min(y)-dy,max(y)+dy,length=ny)}
else ny <- length(ylist)

if (is.null(zlist) )
{dz <- (max(z)-min(z))*extend
zlist <-seq(min(z)-dz,max(z)+dz,length=nz)}
else nz <- length(zlist)

f <- array(0, c(nx,ny,nz))

kx <- matrix(0, n, nx)
for (i in 1:nx) kx[,i]<-dnorm(x,xlist[i],hx*adjust[1])

ky <- matrix(0, n, ny)
for (j in 1:ny) ky[,j]<-dnorm(y,ylist[j],hy*adjust[2])

kz <- matrix(0, n, nz)
for (k in 1:nz) kz[,k]<-dnorm(z,zlist[k],hz*adjust[3])

for (i in 1:nx) for (j in 1:ny) for (k in 1:nz)
f[i,j,k] <- mean(kx[,i]*ky[,j]*kz[,k])

maxf <- max(f)
```

```
levs <- maxf*c(.99,.9,.75,.5,.25,.1)

plot3d(x,y,z,type='n')
contour3d(f, levs, xlist, ylist, zlist,
color=c("black", "red", "orange", "yellow", "cyan", "blue"),
alpha=c(1, 1, .4, .4, .25, .15), add=T, ...)
#decorate3d()

invisible()
}
```

Example 5:
Let us apply tkde to a subset of the states data.

```
library(rgl)
library(misc3d)

tkde(illiteracy, grad, murder)
```

Example 6:
Let us apply tkde to data from a multivariate normal distribution.

```
# 3 independent standard normals

mnorm1 <- matrix(rnorm(3000), ncol = 3)

tkde(mnorm1)

tkde(mnorm1, adjust = c(1.5, 1.5, 1.5))

tkde(mnorm1, adjust = c(2, 2, 2))


## dependent normals

library(MASS)
Sigma <- matrix(c(1,.8,.8,.8,1,.8,.8,.8,1),3,3)     # Covariance matrix
Sigma

mnorm2 <- mvrnorm(1000, mu=rep(0,3), Sigma = Sigma) # MASS library

tkde(mnorm2)

tkde(mnorm2, adjust = c(1.5, 1.5, 1.5))
```

Example 7:

Let us work with the ethanol data from the lattice library.

The R help page for ethanol indicates:

> "Engine exhaust fumes from burning ethanol. Ethanol fuel was burned in a single–cylinder engine. For various settings of the engine compression and equivalence ratio, the emissions of nitrogen oxides were recorded. A data frame with 88 observations on the following 3 variables.
>
> NOx: Concentration of nitrogen oxides (NO and NO2) in micrograms/J.
>
> C: Compression ratio of the engine.
>
> E: Equivalence ratio — a measure of the richness of the air and ethanol fuel mixture.

```
library(lattice)

data(ethanol)
head(ethanol)

compress<-ethanol$C
equiv<-ethanol$E
NOx<-ethanol$NOx

plot3d(compress, equiv, NOx, type = "s", radius = 0.15, col="red")
```

Execute this code once to create the trilpr and trillr functions for local polynomial regression estimates, written by Mike Minnotte:

```
#####################################################################
#                                                                   #
# Functions for calculating and plotting trivariate local           #
# polynomial regression estimates.                                  #
#                                                                   #
# Save as text, then source into R                                  #
#                                                                   #
# trilpr - local polynomial regression of arbitrary order           #
# trillr - local linear regression, much faster than trilpr         #
#                                                                   #
#####################################################################

trilpr<-function(x,y,z,xh,yh,p=1,xlist=NULL,ylist=NULL,extendx=.1,nx=51,
extendy=.1,ny=51,doplot=T,xlab='x',ylab='y',zlab='m(x,y)',phi=30,...)

{# trivariate local polynomial regression (loess) estimate
# x, y - data (explanatory variables)
# z - response variable
# xh, yh - bandwidth (smoothing parameters; standard deviation of normal kernel)
# p - order of locally fitted polynomials (default - linear)
# xlist, ylist - points of evaluation
# extendx, extendy - if xlist (ylist) NULL, how far beyond the range of
#   the data to calculate estimate
# nx, ny - number of points of evaluation
# doplot - if T, plot perspective plot of results, else return (x,y) list for
#   later plotting
# xlab, ylab, zlab - axis labels
if (length(x)!=length(y) | length(x)!=length(z))
stop("Lengths of x, y, and z must be equal")
if (is.null(xlist))
{xr<-range(x)
xd<-(xr[2]-xr[1])*extendx
xlist<-seq(xr[1]-xd,xr[2]+xd,length=nx)}
else
nx<-length(xlist)
if (is.null(ylist))
{yr<-range(y)
yd<-(yr[2]-yr[1])*extendy
ylist<-seq(yr[1]-yd,yr[2]+yd,length=ny)}
else
ny<-length(ylist)

par(err=-1)
mhat<-matrix(0,nx,ny)
n<-length(x)
Y<-matrix(z,n,1)

for (i in 1:nx) for (j in 1:ny)
{W<-dnorm(x,xlist[i],xh)*dnorm(y,ylist[j],yh)
X<-matrix(rep(1,n),ncol=1)
if (p > 0) for (k in 1:p) for (m in 0:k)
X<-cbind(X,(x-xlist[i])^(k-m)*(y-ylist[j])^m)
```

```
betas<-lsfit(X,Y,wt=W,intercept=F)$coef
mhat[i,j]<-betas[1]}

if (doplot) {
persp(xlist,ylist,mhat,phi=phi,...)
return()}
else {
fitted.values<-rep(0,n)
residuals<-rep(0,n)
for (i in 1:n)
{W<-dnorm(x,x[i],xh)*dnorm(y,y[i],yh)
X<-matrix(rep(1,n),ncol=1)
if (p > 0) for (k in 1:p) for (m in 0:k)
X<-cbind(X,(x-x[i])^(k-m)*(y-y[i])^m)
betas<-lsfit(X,Y,wt=W,intercept=F)$coef
fitted.values[i]<-betas[1]
residuals[i]<-z[i]-fitted.values[i]}
return(list(x=xlist,y=ylist,z=mhat,fitted.values=fitted.values,
residuals=residuals))}
}

#########################################################################

trillr<-function(x,y,z,xh,yh,xlist=NULL,ylist=NULL,extendx=.1,nx=51,
extendy=.1,ny=51,doplot=T,xlab='x',ylab='y',zlab='m(x,y)',phi=30,...)

{# trivariate local linear regression (loess) estimate
# x, y - data (explanatory variables)
# z - response variable
# xh, yh - bandwidth (smoothing parameters; standard deviation of normal kernel)
# xlist, ylist - points of evaluation
# extendx, extendy - if xlist (ylist) NULL, how far beyond the range of
#   the data to calculate estimate
# nx, ny - number of points of evaluation
# doplot - if T, plot perspective plot of results, else return (x,y) list for
#   later plotting
# xlab, ylab, zlab - axis labels
if (length(x)!=length(y) | length(x)!=length(z))
stop("Lengths of x, y, and z must be equal")
if (is.null(xlist))
{xr<-range(x)
xd<-(xr[2]-xr[1])*extendx
xlist<-seq(xr[1]-xd,xr[2]+xd,length=nx)}
else
nx<-length(xlist)
if (is.null(ylist))
{yr<-range(y)
yd<-(yr[2]-yr[1])*extendy
ylist<-seq(yr[1]-yd,yr[2]+yd,length=ny)}
else
ny<-length(ylist)

par(err=-1)
mhat<-matrix(0,nx,ny)
```

```
n<-length(x)
Y<-matrix(z,n,1)

for (i in 1:nx)
{Kx<-dnorm(x,xlist[i],xh)
dx<-x-xlist[i]
dx2<-dx^2
for (j in 1:ny)
{K<-Kx*dnorm(y,ylist[j],yh)
dy<-y-ylist[j]
T0<-sum(K)
Tx<-sum(K*dx)
Txx<-sum(K*dx2)
Ty<-sum(K*dy)
Tyy<-sum(K*dy^2)
Txy<-sum(K*dx*dy)
Tz<-sum(K*z)
Txz<-sum(K*dx*z)
Tyz<-sum(K*dy*z)
Sxxyy<-Txx*Tyy-Txy^2
Sxyy<-Ty*Txy-Tx*Tyy
Sxxy<-Tx*Txy-Ty*Txx
mhat[i,j]<-(Tz*Sxxyy+Txz*Sxyy+Tyz*Sxxy)/
(T0*Sxxyy+Tx*Sxyy+Ty*Sxxy)}}
if (doplot) {
persp(xlist,ylist,mhat,phi=phi,...)
return()}
else {
fitted.values<-rep(0,n)
residuals<-rep(0,n)
for (i in 1:n)
{K<-dnorm(x,x[i],xh)*dnorm(y,y[i],yh)
dx<-x-x[i]
dy<-y-y[i]
T0<-sum(K)
Tx<-sum(K*dx)
Txx<-sum(K*dx^2)
Ty<-sum(K*dy)
Tyy<-sum(K*dy^2)
Txy<-sum(K*dx*dy)
Tz<-sum(K*z)
Txz<-sum(K*dx*z)
Tyz<-sum(K*dy*z)
Sxxyy<-Txx*Tyy-Txy^2
Sxyy<-Ty*Txy-Tx*Tyy
Sxxy<-Tx*Txy-Ty*Txx
fitted.values[i]<-(Tz*Sxxyy+Txz*Sxyy+Tyz*Sxxy)/
(T0*Sxxyy+Tx*Sxyy+Ty*Sxxy)
residuals[i]<-z[i]-fitted.values[i]}
return(list(x=xlist,y=ylist,z=mhat,fitted.values=fitted.values,
residuals=residuals))}
}
```

```
NOx.fit <- trillr(compress, equiv, NOx, 1, 0.07, doplot = F)
persp(NOx.fit, phi = 30, theta = -70)

persp3d(NOx.fit, col = "grey", xlab = "C", ylab = "E", zlab = "NOx")
spheres3d(compress, equiv, NOx, type = "s", radius = 0.15, col = "red")

persp3d(NOx.fit, col = "grey", xlab = "C", ylab = "E", zlab = "NOx", alpha = 0.8)
spheres3d(compress, equiv, NOx, type = "s", radius = 0.15, col = "red")
```

8 "Hypervariate" (High–Dimensional) Plots

8.1 Scatterplot Matrix (for $n \geq 4$)

Example 1:

```
data(state)
head(state.x77)

pairs(state.x77)

pairs(state.x77[,2:7])
```

Example 2:

```
data(iris)
head(iris)

species <- iris[,5]
iris <- iris[,1:4]

pairs(iris)

pairs(iris, pch = 21, bg = unclass(species))
```

Lecture 31:
Fr 03/27/09

8.2 Parallel Coordinate Plots

Symanzik (2004), p. 303, states:

> "Parallel coordinate plots (Inselberg 1985, Wegman 1990) [...] are a geometric device for displaying points in high–dimensional spaces, in particular, for dimensions greater than three. The idea is to sacrifice orthogonal axes by drawing the axes parallel to each other resulting in a planar diagram where each d–dimensional point $(x_1, \ldots, x_d)$ is uniquely represented by a continuous line. The parallel coordinate representation enjoys some elegant duality properties with the usual Cartesian coordinates and allows interpretations of statistical data in a manner quite analogous to two–dimensional Cartesian scatterplots. This duality of lines in Cartesian plots and points in parallel coordinates extends to conic sections. This means that an ellipse in Cartesian coordinates maps into a hyperbola in parallel coordinates. Similarly, rotations in Cartesian coordinates become translations in parallel coordinates.
>
> The individual parallel coordinate axes represent one–dimensional projections of the data. We can isolate clusters by looking for separation between data points on any axis or between any pair of axes. Because of the connectedness of the multidimensional parallel coordinate diagram, it is usually easy to see whether or not this clustering propagates through other dimensions."

Example 3:

```
library(MASS)

state<-state.x77[,2:7]
state2<-state
head(state2)

state2[,c(2,4,6)] <- -state2[,c(2,4,6)] #large = good for all vars
head(state2)
```

```
parcoord(state)

parcoord(state2)

parcoord(state2, col = unclass(state.region))
```

Example 4:

```
parcoord(iris) # produces strange warning

parcoord(iris, pch = rep(1,4))

parcoord(iris, col = unclass(species), pch = rep(1, 4))
```

Example 5:

```
library (colorspace)

x = seq(0, 2 * pi, length.out = 30)
sinx = sin(x)
cosx = cos(x)

par (mfrow = c(1, 2))

plot(sinx, cosx, pch = 22, col = topo.colors(30))

parcoord(cbind(sinx, cosx), col = topo.colors(30),
  var.label = TRUE, pch = rep(1, 2))
```

8.3 Faces, Star Plots, and other Glyph Representations

Venables & Ripley (2002), p. 314, state:

> "There is a wide range of ways to trigger multiple perceptions of a figure, and we can use these to represent each of a moderately large number of rows of a data matrix by an individual figure. Perhaps the best known of these are Chernoff's faces (Chernoff 1973) [...] and the star plots as implemented in the function stars [...].
>
> These glyph plots do depend on the ordering of variables and perhaps also their scaling, and they do rely on properties of human visual perception. So they have rightly been criticised as subject to manipulation, and one should be aware of the possibility that the effect may differ by viewer. Nevertheless they can be very effective as tools for private exploration."

Other glyph representations discussed in the literature are based on "trees" and "castles" (Kleiner & Hartigan 1981).

Example 6:

```
data(state)

state <- state.x77[,2:7]
state2 <- state
state2[,c(2,4,6)] <- -state2[,c(2,4,6)] #large = good for all vars

stars(state, key.loc = c(15, 1.5))  #star plot of raw state data

stars(state2, key.loc = c(15, 1.5))
```

Example 7:

```
data(iris)

species <- iris[,5]
iris <- iris[,1:4]

stars(iris, key.loc = c(25, 1.3))

stars(iris, key.loc = c(25, 1.3), col.stars = unclass(species))
```

The R help page for faces indicates:

> "Chernoff Faces: [...] Explanation of parameters: 1-height of face, 2-width of face, 3-shape of face, 4-height of mouth, 5-width of mouth, 6-curve of smile, 7-height of eyes, 8-width of eyes, 9-height of hair, 10-width of hair, 11-styling of hair, 12-height of nose, 13-width of nose, 14-width of ears, 15-height of ears. For details look at the literate program of faces."

Example 8:

```
library(aplpack)

faces(state2)

# look at different permutations of the variable order
faces(state2[,c(2:6, 1)])

faces(state2[,c(3:6, 1:2)])
```

Example 9:

```
data(iris)

species <- iris[,5]
iris <- iris[,1:4]

faces(iris)
```

You should decide yourself how useful these are!

8.4 Andrews Plots

Symanzik (2004), p. 306, states:

> "The Andrews (multidimensional data) plot, as introduced in Andrews (1972) is based on a series of Fourier interpolations of the coordinates of multi–dimensional data points. Points that are close in some metric will tend to have similar Fourier interpolations and therefore will tend to cluster in the Andrews plot. Thus, the Andrews plot is an informative graphical tool most useful to detect clustering.
>
> Ideas underlying the Andrews plot and the grand tour are quite similar. However, in contrast to the grand tour, the Andrews plot is a static plot while the grand tour is dynamic. Although dynamic renditions of the Andrews plot exist, and these sometimes also are (incorrectly) referred to as one–dimensional grand tour (Crawford & Fall 1990), the Andrews plot is not a grand tour since it cannot sweep out all possible directions as pointed out in Wegman & Shen (1993). Three–dimensional generalizations of the Andrews plot and other pseudo grand tours have been introduced in Wegman & Shen (1993) as well."

Execute this code once to create the Andrews function, written by Mike Minnotte:

```
#####################################################################
#                                                                   #
# Function for calculating and plotting Andrews curves for          #
# visualization of multivariate data                                #
#                                                                   #
# Save as text, then source into R                                  #
#                                                                   #
#####################################################################


Andrews<-function(x,ord=1:ncol(x),nt=101,col=1,lty=1,...)


{# calculates and plots Andrews curve representations of multivariate data
# x - matrix of data, rows are observations, columns are variables
# ord - order of variables for Fourier representation.  Earlier variables
#    are lower frequency terms.
# nt - number of horizontal points to plot at (default 101).
x<-x[,ord]
minx<-apply(x,2,min)
maxx<-apply(x,2,max)
rangex<-maxx-minx
for (i in 1:ncol(x))
x[,i]<-(x[,i]-minx[i])/rangex[i]
x<-2*x-1
t<-seq(-pi,pi,length=nt)
nx<-nrow(x)
Aout<-matrix(0,nx,nt)
for (i in 1:nx)
Aout[i,]<-rep(x[i,1]/sqrt(2),nt)

np<-floor((ncol(x)-1)/2)

if (np > 0)
for (j in 1:np)
for (i in 1:nx)
Aout[i,]<-Aout[i,]+x[i,2*j]*sin(t*j)+x[i,2*j+1]*cos(t*j)

if (ncol(x) > 2*np+1)
for (i in 1:nx)
Aout[i,]<-Aout[i,]+x[i,ncol(x)]*sin(t*(np+1))

#plot(t,Aout[1,],type='l',ylim=range(Aout))
#for (i in 2:nx)
# lines(t,Aout[i,])

matplot(t,t(Aout),type='l',xlab='t',ylab='A(t)',main='Andrews Curves',
col=col,lty=lty,...)

return()
}
```

Example 10:

```
data(state)

state <- state.x77[,2:7]

Andrews(state)

Andrews(state, ord = 6:1)

Andrews(state, ord = 6:1, col = unclass(state.region))
legend("bottomleft", levels(state.region), lty = 1, col = 1:4)
```

Example 11:

```
data(iris)

species <- iris[,5]
iris <- iris[,1:4]

Andrews(iris, col = unclass(species))
legend("bottomleft", levels(species), lty = 1, col = 1:3)

Andrews(iris, ord = 4:1, col = unclass(species))
legend("bottomleft", levels(species), lty = 1, col = 1:3)
```

8.5 Data Images

Minnotte & West (1998), p. 25, state:

> "The color histogram was first introduced as a tool for visualizing higher dimensional data by Wegman (1990). A new version of this concept, called a data image, is discussed. Each variable is transformed into a greyscale or color range so that a high–dimensional data set may be viewed as an image, with observations on one axis and variables on the other. The rows and columns of the image may be rearranged to highlight relationships between observations and variables. New ways of displaying the image based on various linear orderings of a data set are discussed."

A Java version of the data image software, developed by Mike Minnotte and Webster West, can be accessed at `http://www.math.usu.edu/~minnotte/research/java/dimage.html`.

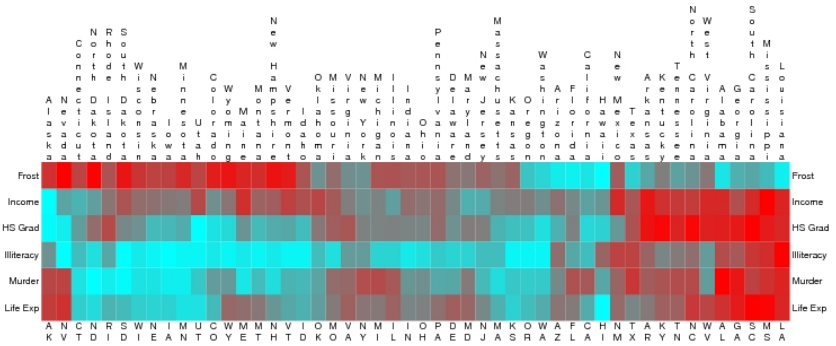


Figure 6: Data image of 1970's states data. Cyan is high for income, high school graduation, and life expectancy; low for frost days, illiteracy rate, and murder rate. Both observations and variables are sorted by complete linkage clustering algorithm. Identification of observations as well as variables allows recognition of expected and surprising clusters.

Figure 89: Minnotte & West (1998), p. 30, Figure 6: Screenshot of color version taken from `http://www.math.usu.edu/~minnotte/research/preprints/DataImage-col.ps` on 3/26/2009.

Example 12:

Source the R code once to create the dataimage function, written by Mike Minnotte, then execute the remaining code:

```
source(url("http://www.math.usu.edu/~symanzik/teaching/2009_stat6560/RDataAndScripts/Ch8_Dataimage.R"))

library(stats)

data(state)
state <- state.x77[,2:7]
state2 <- state
state2[,c(2,4,6)] <- -state2[,c(2,4,6)] #large = good for all vars

dataimage(state2)
```

Example 13:

```
data(iris)

iris <- as.matrix(iris[,1:4])

dataimage(iris)
```

9 Statistical Maps

9.1 Choropleth Maps

Symanzik & Carr (2008), p. 270, state:

> "Choropleth maps use the color or shading of regions in a map to represent region values. Choropleth maps have proved very popular but have many problems and limitations as indicated by writers such as Robinson et al. (1978), Dent (1993), and Harris (1999). [...]
>
> There are two kinds of choropleth maps, called unclassed and classed. Unclassed maps use a continuous color scale to encode continuous values (statistics). This is problematic because perception of color is relative to neighboring colors and because color has poor perceptual accuracy of extraction in a continuous context. Classed choropleth maps ameliorate this problem and dominate in the literature.
>
> Classed choropleth maps use class intervals to convert continuous estimates into an ordered variable with a few values that can be represented using a few colors. When a few colors are easily discriminated and regions are sufficiently large for color perception, color identification problems are minimal. The color scheme also needs to convey the class ordering based on values. Brewer (1997) and Brewer et al. (1997) provided results evaluating different color schemes in a mapping context. The Web site `http://colorbrewer.org` (see Leslie 2002, for a short description) contains guidance on ordered color schemes and additional issues such as suitable schemes for people with color vision deficiencies and for different media. Perfect examples on how colors should be used in choropleth maps can be found in the 1996 "Atlas of United States Mortality" (Pickle et al. 1996)."

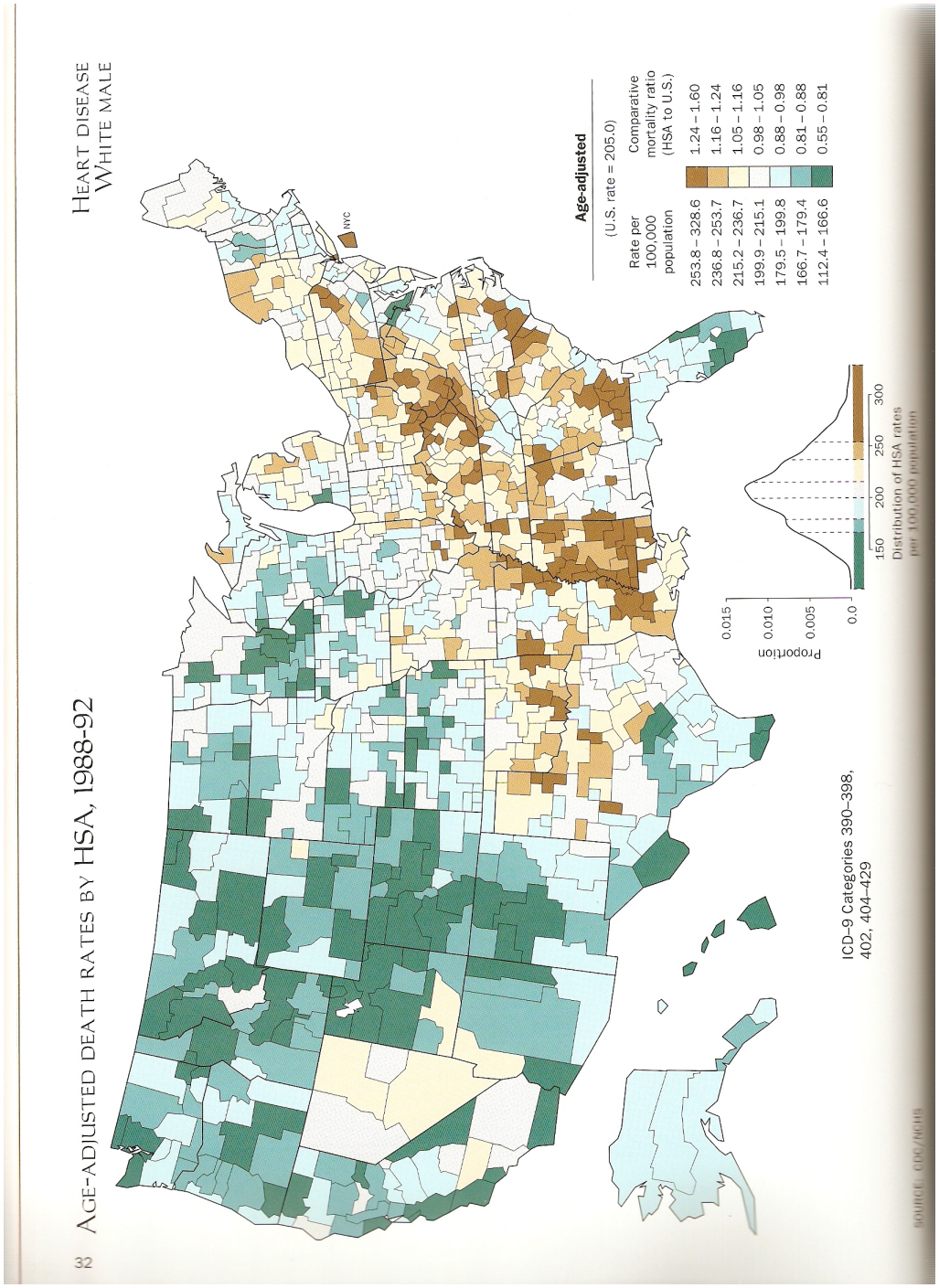


Figure 90: Pickle et al. (1996), p. 32, Figure, showing Heart Disease White Male by Health Service Area (HSA), 1988–92.

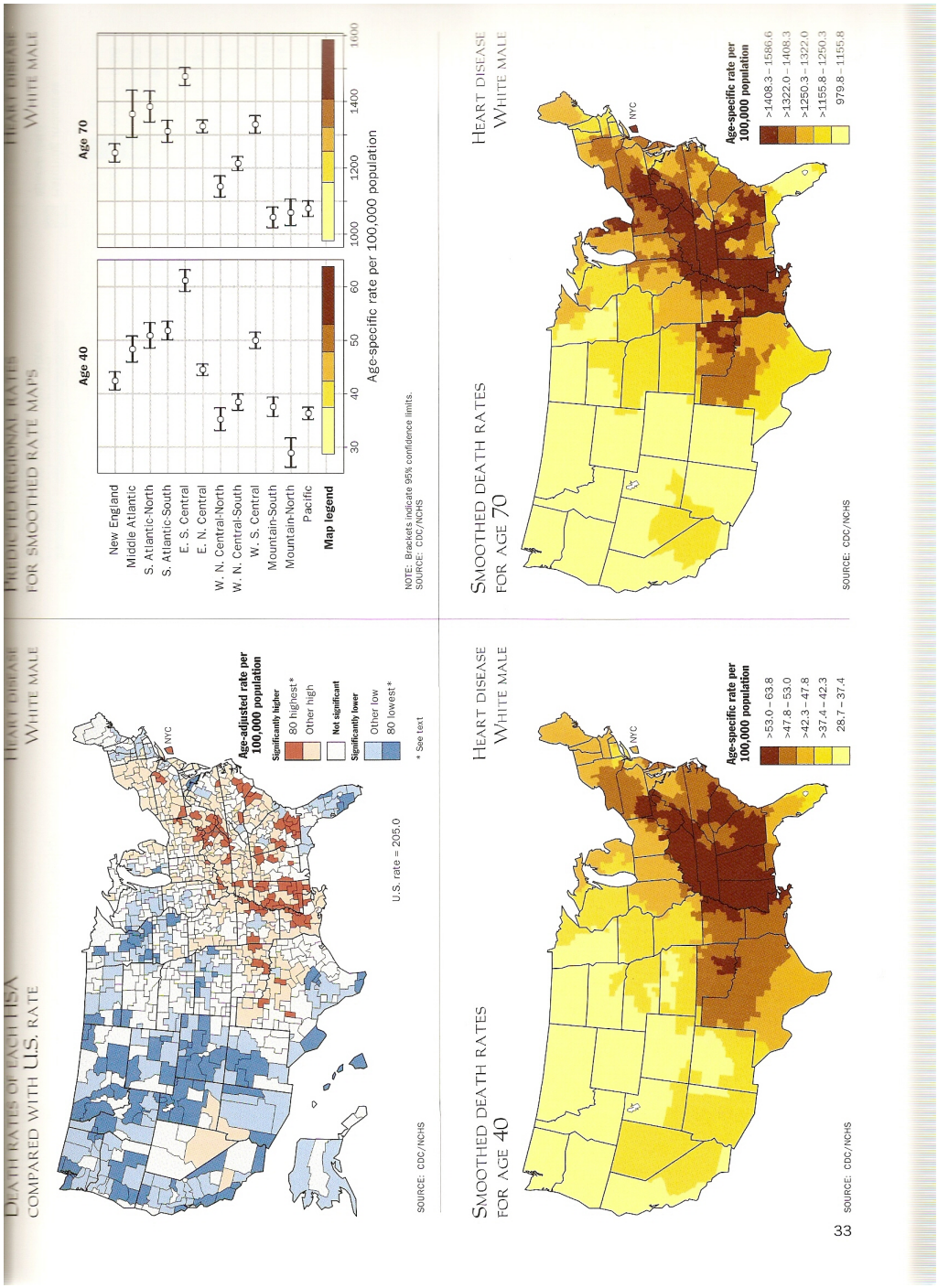


Figure 91: Pickle et al. (1996), p. 33, Additional Figures, showing Heart Disease White Male, 1988–92.

Choropleth Map Construction:

Choropleth Maps are easy to construct on paper. `http://www.teachingideas.co.uk/geography/chlormap.htm` teaches children age 5 to 11 "How to make a Choropleth Map".

Weather Maps:

We are used to see various new choropleth maps almost daily, in newspapers, on TV, or on the Web. Probably the most popular choropleth maps are weather maps, such as provided by `http://www.usatoday.com/weather/default.htm`. Move your mouse over the various options (Radar, Satellite, Precip, Temps, Fronts) to see various weather–related choropleth maps.

Unfortunately, except for Temps, there are usually no exact values shown on choropleth maps. So, it is difficult to exactly compare different regions or cities.

9.2 Choropleth Maps in R

Various R packages support the creation of maps and choropleth maps in R, provide commands to read and write ESRI shapefiles, and link to various Geographic Information Systems (GIS). A more detailed overview can be found at `http://geodacenter.asu.edu/map-packages-on-cran`.

Example 1:
Basic geographic maps, ranging from a world map to detailed county maps with labels and city names.

```
library(maps)

map() # World (default)

map("state") # US

map("state", c("Utah", "Colorado", "Idaho", "Wyoming", "Montana"))

data(state)
state.region
state.name[state.region == "Northeast"]
map("state", state.name[state.region == "Northeast"])

map("state")
map.axes() # add latitude (North/South) and longitude (East/West)

map("county") # US counties

map("county", "Utah") # Utah counties

map.text("state") # too many labels

map.text("state", state.name[state.region == "South"])

map("county", "Utah")
map.cities() # too many labels
```

```
map("county", "Utah", xlim = c(-113,-111), ylim = c(40,41))
map.cities()
map.axes()

map("state", state.name[state.region == "Northeast"], fill = T, col = 0:8)

## use better colors from RColorBrewer

library(RColorBrewer)

map("state", state.name[state.region == "Northeast"], fill = T,
  col = brewer.pal(9, "Set3"))

map("state", fill = T, col = brewer.pal(9, "Set3"))

map("state", fill = T, col = brewer.pal(5, "Blues"))
```

Example 2:
Choropleth maps.

```
data(state)

murder <- state.x77[,5]

range(murder)

breaks <- c(1, 4, 7, 10, 13, 16)
m.class <- cut(murder, breaks)

brewer.pal(5, "Blues")
m.col <- brewer.pal(5, "Blues")[m.class]

map.m.col <- m.col[match.map("state", state.name)]
map("state", fill = T, col = map.m.col)
legend("bottomright", legend = levels(m.class), fill = brewer.pal(5, "Blues"))
```

9.3 Micromaps

Symanzik & Carr (2008), p. 268, state:

> "Over the last decade, researchers have developed many improvements to make statistical graphics more accessible to the general public. These improvements include making statistical summaries more visual and providing more information at the same time. Research in this area involved converting statistical tables into plots (Carr 1994, Carr & Nusser 1995), new ways of displaying geographically referenced data (Carr et al. 1992), and, in particular, the development of linked micromap (LM) plots, often simply called micromaps (Carr & Pierson 1996, Carr et al. 1998, Carr, Olsen, Pierson & Courbois 2000). LM plots, initially called map row plots as well as linked map–attribute graphics, were first presented in a poster session sponsored by the American Statistical Association (ASA) Section on Statistical Graphics at the 1996 Joint Statistical Meetings (Olsen et al. 1996). More details on the history of LM plots and their connection to other research can be found in these early references on micromaps. More recent references on LM plots (Carr, Wallin & Carr 2000, Carr 2001) focused on their use for communicating summary data from health and environmental studies."

9.3.1 Template for LM Plots

Gebreab et al. (2008), pp. 112–113, state:

> "A typical template of a LM plot consists of four key features (Carr & Pierson 1996). Figure 1 shows a hypothetical LM plot. The first feature is **three or more sequence panels in parallel linked by location**. In the hypothetical case, Figure 1 shows five parallel sequences of panels. The first (leftmost) sequence of panels is the micromap panel itself that typically contains small caricatures of map outlines of a region. The caricature map maintains the shape and neighborhood relationship while making the small subregions more visible. The second (from the left) sequence of panels is the label panel that provides the names of the geographical subregions (here, Region 1 through Region 10). The third through the fifth (from the left) sequence of panels display the statistical summaries. These panels may represent many forms of statistical summaries including box–plots, dot–plots (as shown in Figure 1), time series plots, confidence intervals, etc. **Sorting the geographic subregions based on the statistical variable(s) of interest** is the second feature. Sorting improves perception between consecutive panels from the top to the bottom of the display. The third feature is the **partitioning of the regions into perceptual groups of size five or less to allow the viewer's attention to focus on explicit areas at a time**. The fourth feature is **color and location that links corresponding elements within the parallel sequence panels**, i.e., the color red in the topmost panels relates to the geographic subregion in the northeast of the map, the subregion name (Region 5), and a red dot in each of the three statistical panels. The color red is reused in the next consecutive set of panels for Region 2, but there is no relationship between Region 5 and Region 2 as one might at first assume. Simply, there do not exist enough distinguishable colors to populate an entire display (with, say, 50 different subregions) such that colors have to be reused in different panels.
>
> In the hypothetical Figure 1, the rows are sorted by decreasing values with respect to the statistical panel 2. The statistical data displayed in the statistical panel 1 and 2 show a strong positive association (the correlation r calculated as 0.99), expressed in the almost parallel behavior of the dots and lines representing the values for these two variables. In contrast, the statistical data in panel 3 and 1 (as well as 3 and 2) show a strong negative

association (the correlation r calculated as Ű0.94 for 3 and 1 and as Ű0.92 for 3 and 2). This negative association is seen in the movement of the dots and lines in opposite directions for these variables. Moreover, the data in panel 3 shows an unusual outlier, the value for Region 1. It is this outlier that considerably reduces the almost perfect negative association otherwise present in this data. Just a simple numerical calculation of r might not be able to reveal the influence of a single subregion on the overall relationship.

The map panels of the LM plot in Figure 1 exhibit a strong geographic pattern: Highest occurrences with respect to the statistical panel 1 and 2 can be found in the north and in the east; lowest occurrences can be found in the west and in the south. Additional features of LM plots exist and are described in more details in Symanzik & Carr (2008)."

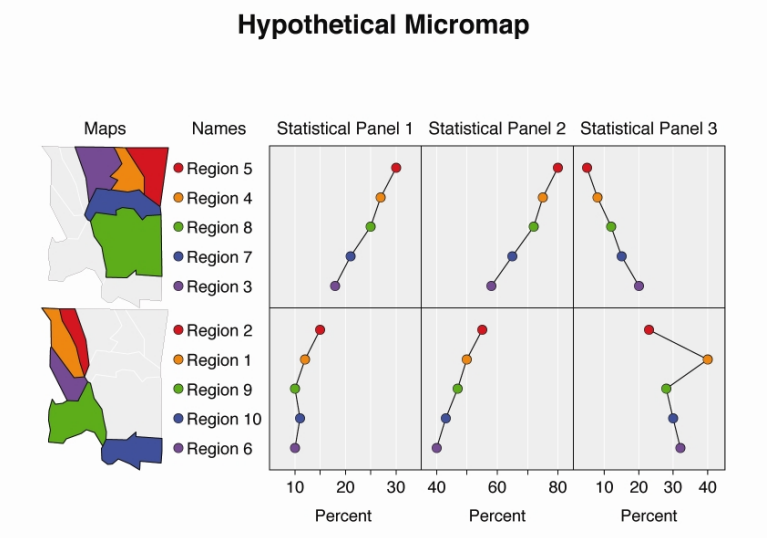


Figure 92: Gebreab et al. (2008), p. 113, Figure 1: Hypothetical LM plot illustrating the main features of such plots: the leftmost sequence of map panels, the second (from the left) sequence of label panels, and the third through the fifth (from the left) sequence of statistical panels.

Lecture 34:
We 04/03/09

9.3.2 Micromaps vs. Choropleth Maps

Symanzik & Carr (2008), pp. 270–271, state:

> "LM plots often provide a good alternative to displaying statistical information using choropleth maps. Choropleth maps use the color or shading of regions in a map to represent region values. Choropleth maps have proved very popular but have many problems and limitations as indicated by writers such as Robinson et al. (1978), Dent (1993), and Harris (1999). Reviewing these problems helps to indicate why LM plots are a good alternative. [...]
>
> Even with a good color scheme, three key problems remain for classed choropleth maps. **The first problem relates to region area. As suggested above, some map regions can be too small to effectively show color.** Examples include Washington, D.C., on a map of the United States (U.S.) and Luxembourg on a European map. Map caricatures, such as Monmonier's state visibility map (Monmonier 1993), can address this problem, by enlarging small regions in a way that maintains region identifiability and shows each region touching the actual neighboring regions. Another facet of the area problem is that large areas have a strong visual impact while in many situations, such as in the mapping of mortality rates, the interpretation should be weighted by the region population. Dorling (1995) addressed this problem by constructing cartograms that changed region shapes to make areas proportional to population. Issues related to this approach are region identifiability, and map instability over time as their shapes change with changing populations. Area related problems persist in choropleth maps.
>
> **A second key problem is that converting a continuous variable into a variable with a few ordered values results in an immediate loss of information.** This loss includes the relative ranks of regions whose distinct values become encoded with the same value. The task of controlling conversion loss has spawned numerous papers about proposing methods for defining class intervals. Today, guidance is available based on usability studies. Brewer & Pickle (2002) indicated that quintile classes (roughly 20% of regions in each class) tend to perform better than other class interval meth-

ods when evaluated across three different map reading tasks. Still, the best of the class interval selection approaches loses information.

The third key problem is that it is difficult to show more than one variable in a choropleth map. MacEachren et al. (1995) and MacEachren et al. (1998) were able to clearly communicate values of a second binary variable (and indicator of estimate reliability) by plotting black and white stripped texture on regions with uncertain estimates. However, more general attempts such as using bivariate colors schemes have been less successful (Wainer & Francolini 1980). Thus, choropleth maps are not suitable for showing estimate standard errors and confidence bounds that result from the application of sound statistical sampling or description. It is possible to use more than one map to show additional variables. However, Monmonier (1996, page 154) observed that when plotting choropleth maps side by side it can easily happen that "similarity among large areas can distort visual estimates of correlation by masking significant dissimilarity among small areas." The change blindness (Palmer 1999, page 538) that occurs as the human eyes jump from one map to another map makes it difficult to become aware of all the differences that exist in multiple choropleth maps and hard to mentally integrate information in a multivariate context."

Symanzik & Carr (2008), pp. 272–274, provide the following motivational example:

"Fig. 93 shows two variables, the soybean yield and acreage from the 1997 Census of Agriculture for the United States, displayed in two choropleth maps. Five equal size class intervals were chosen for each of the maps. [...]

The two choropleth maps in Fig. 93 indicate that highest yields and highest acreages for soybeans occur in the Midwest. There seems to be some spatial trend, i.e., some steady decrease for both variables from the Midwest to the Southeast. Overall, there appears to be a positive correlation between these two variables since high yields/high acreages and low yields/low acreages seem to appear in the same geographic regions. The correlation coefficient between yield and acreage is only 0.64, suggesting departures from linearity that would be better revealed using scatterplots or LM plots. [...]

In fact, Fig. 94 shows the LM plots of the same two variables as Fig. 93, plus a third statistical panel for the variable production. Data is available for 31 of the 50 U.S. states only. An identical color links all of the descriptors

for a region. Successive perceptual groups use the same set of distinct colors. In Fig. 94, the sorting is done (from largest to smallest) by soybean yield in those 31 U.S. states where soybeans were planted. Here, the points within a panel are connected to guide the viewer's eyes and not to imply that interpolation is permitted. The connecting lines are a design option and can be omitted to avoid controversy or suit personal preferences. The list of 31 U.S. states is not evenly divisible by five. Two perceptual groups at the top and two groups at the bottom contain four states, while three perceptual groups in the middle contain five states. The middle groups require the use of a fifth linking color. Using distinct hues for linking colors works best in full color plots. For grey–level plots, colors need to be distinct in terms of grey–level. Fig. 94 shows different shades of green and is suitable for production as a grey–level plot. Readers new to LM plots sometimes try to compare regions with the same color across the different perceptual groups, but quickly learn the linkage is meaningful only within a perceptual group."

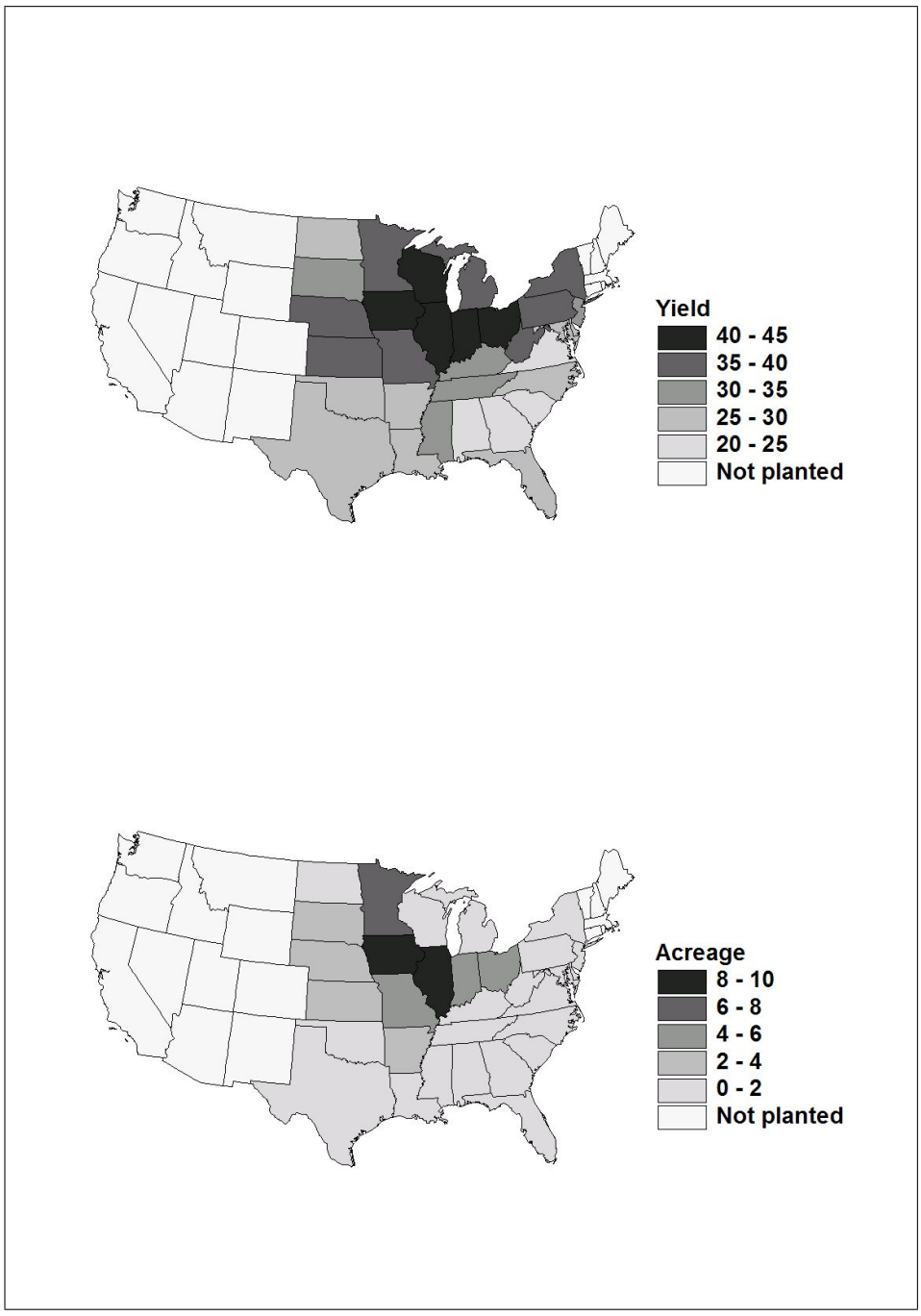


Figure 93: Symanzik & Carr (2008), p. 273, Figure 1.2: Choropleth maps of the 1997 Census of Agriculture, showing the variables soybean yield (in bushels per acre) and acreage (in millions of acres) by state. The data represent the 31 U.S. states where soybeans were planted.

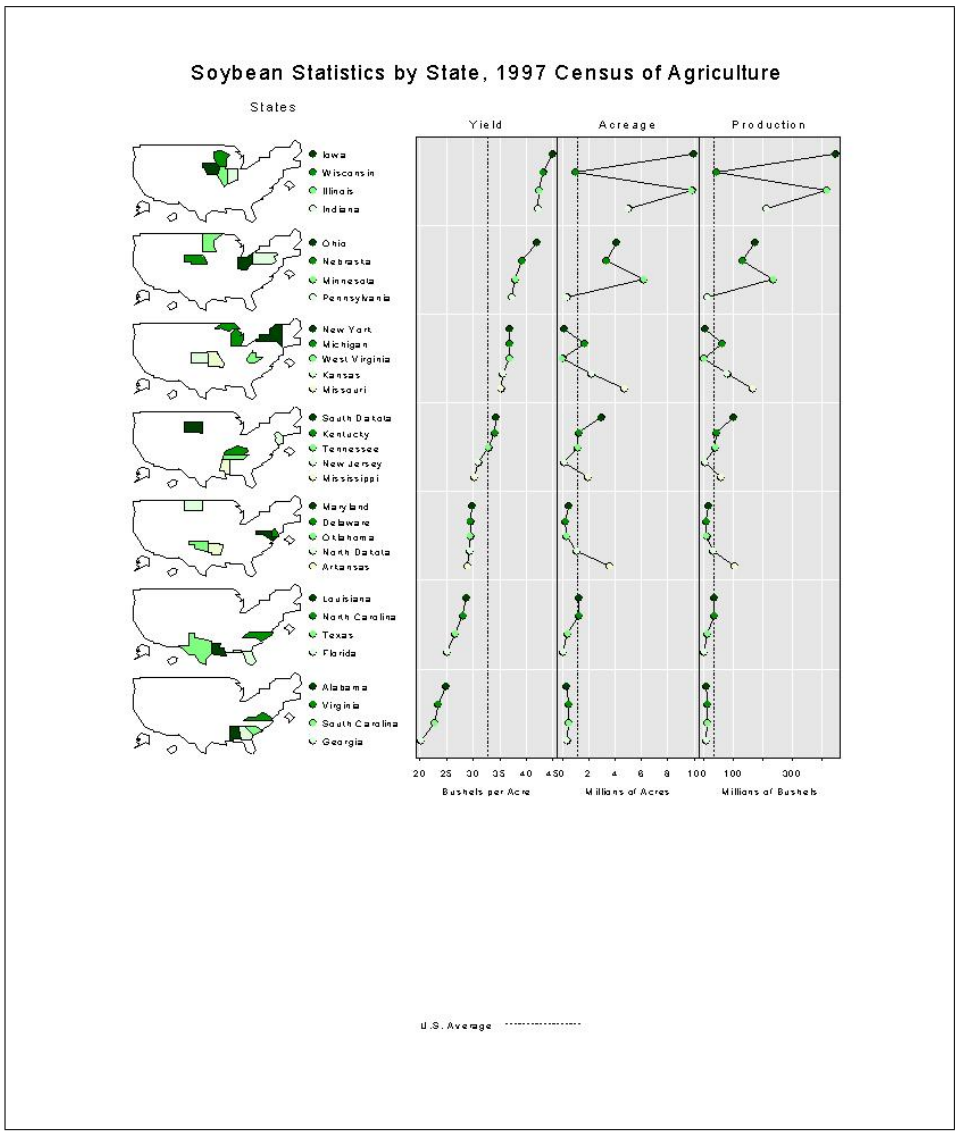


Figure 94: Symanzik & Carr (2008), p. 275, Figure 1.3: LM plots of the 1997 Census of Agriculture, showing soybean yield (in bushels per acre), acreage (in millions of acres), and production (in millions of bushels) by state. The data is sorted by yield and shows the 31 U.S. states where soybeans were planted. The "U.S. Average" represents the median, i.e., the value that splits the data in half such that one half of the states has values below the median and the other half of the states has values above the median. For example, Tennessee is the state with the median yield. This figure has been republished from `http://www.nass.usda.gov/research/gmsoyyap.htm` without any modifications (and ideally should contain much less white space in the lower part).

9.3.3 Additional Micromap Examples

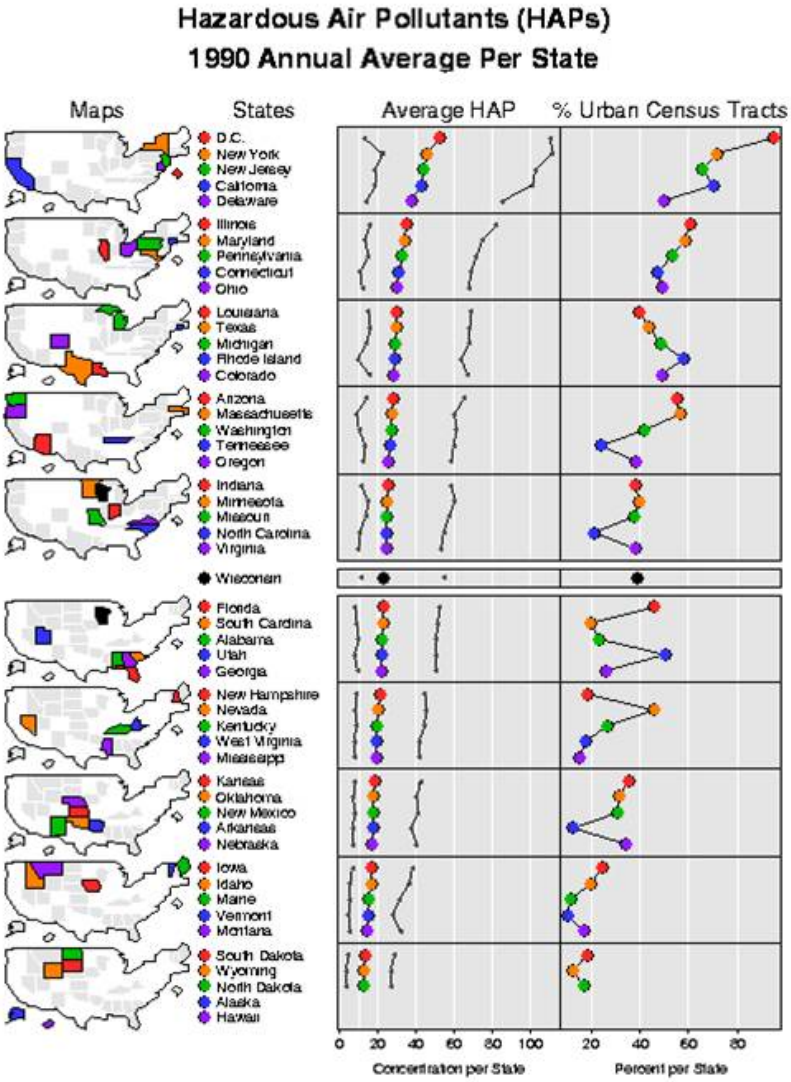


Figure 95: Hazardous Air Pollutants (HAPs) for the US from work with the EPA.

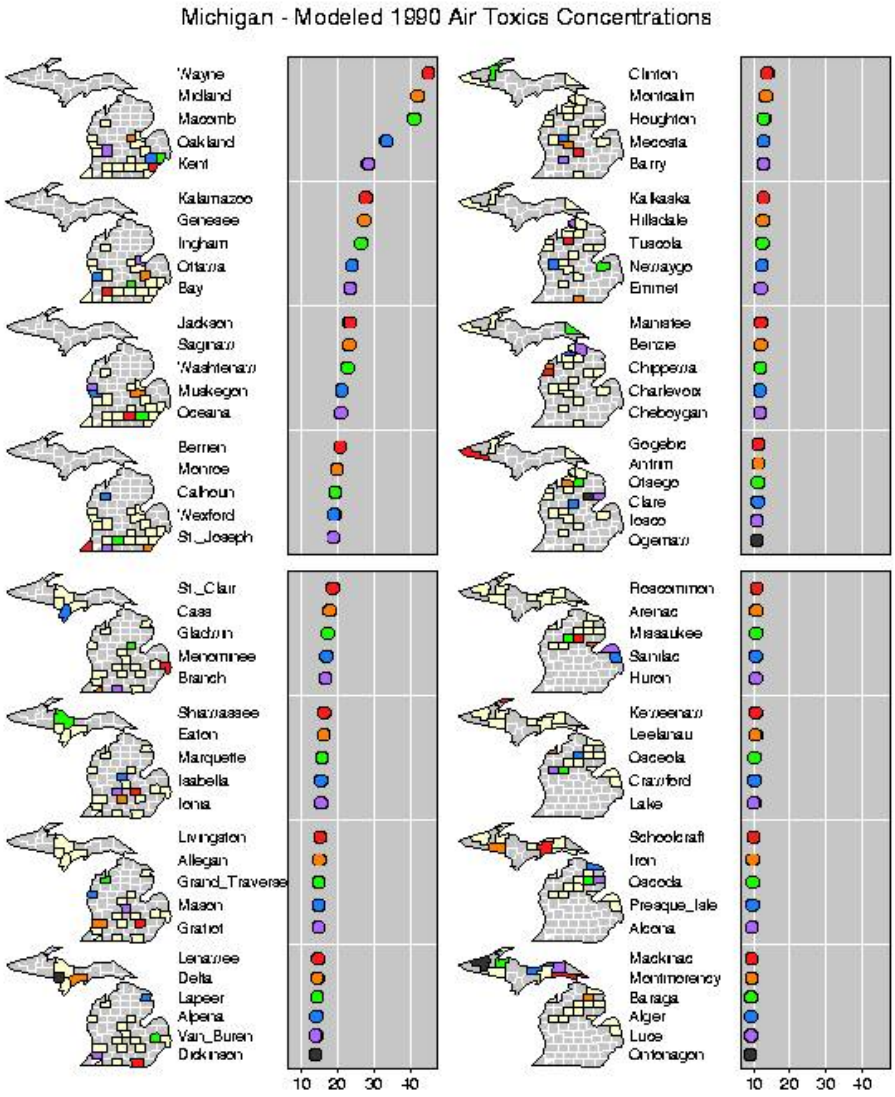


Figure 96: Hazardous Air Pollutants (HAPs) for Michigan from work with the EPA.

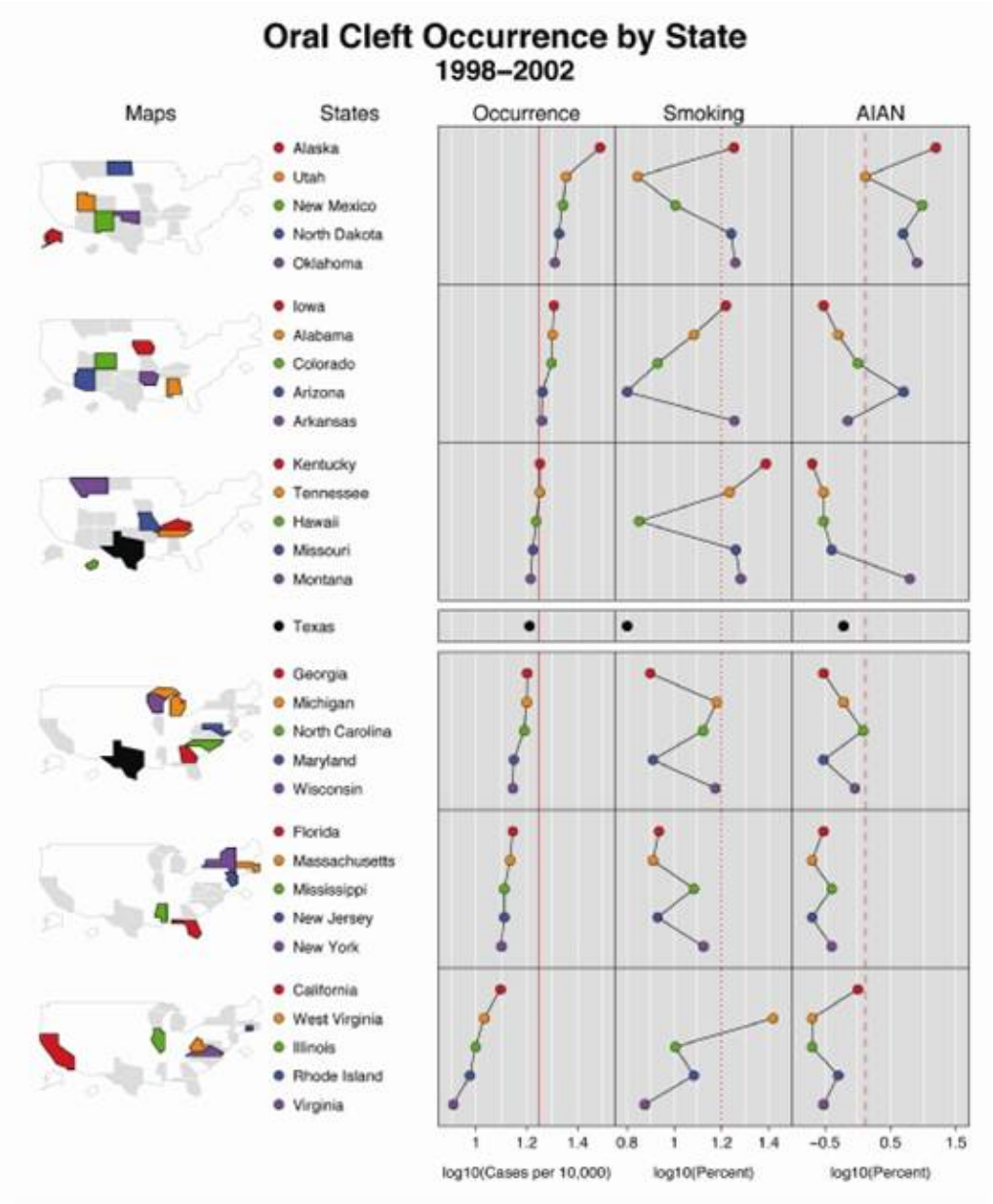


Figure 97: Gebreab et al. (2008), p. 114, Figure 2: LM plot showing oral cleft occurrence by state for the period of 1998–2002. Only oral cleft occurrence for 31 out of the 50 US states was available and displayed here. Smoking rates for California were not available. The red lines show the national average (i.e., mean) of oral cleft occurrence (17.7 per 10,000), smoking rate 16%, and AINA proportion of 1.3%. Note that Texas had the median oral cleft occurrence among the 31 states for which data were available.

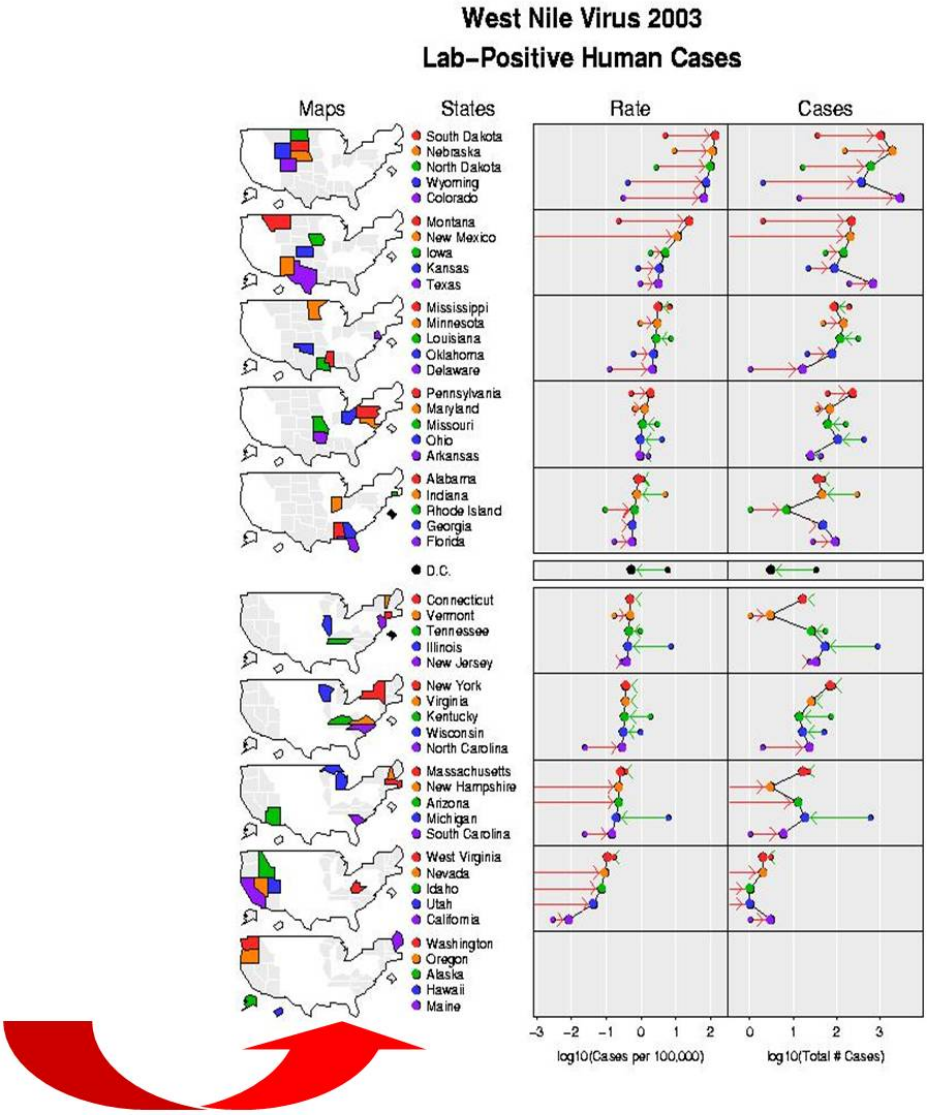


Figure 98: Spread of the West Nile Virus (WNV) across the US, 2002 vs 2003.

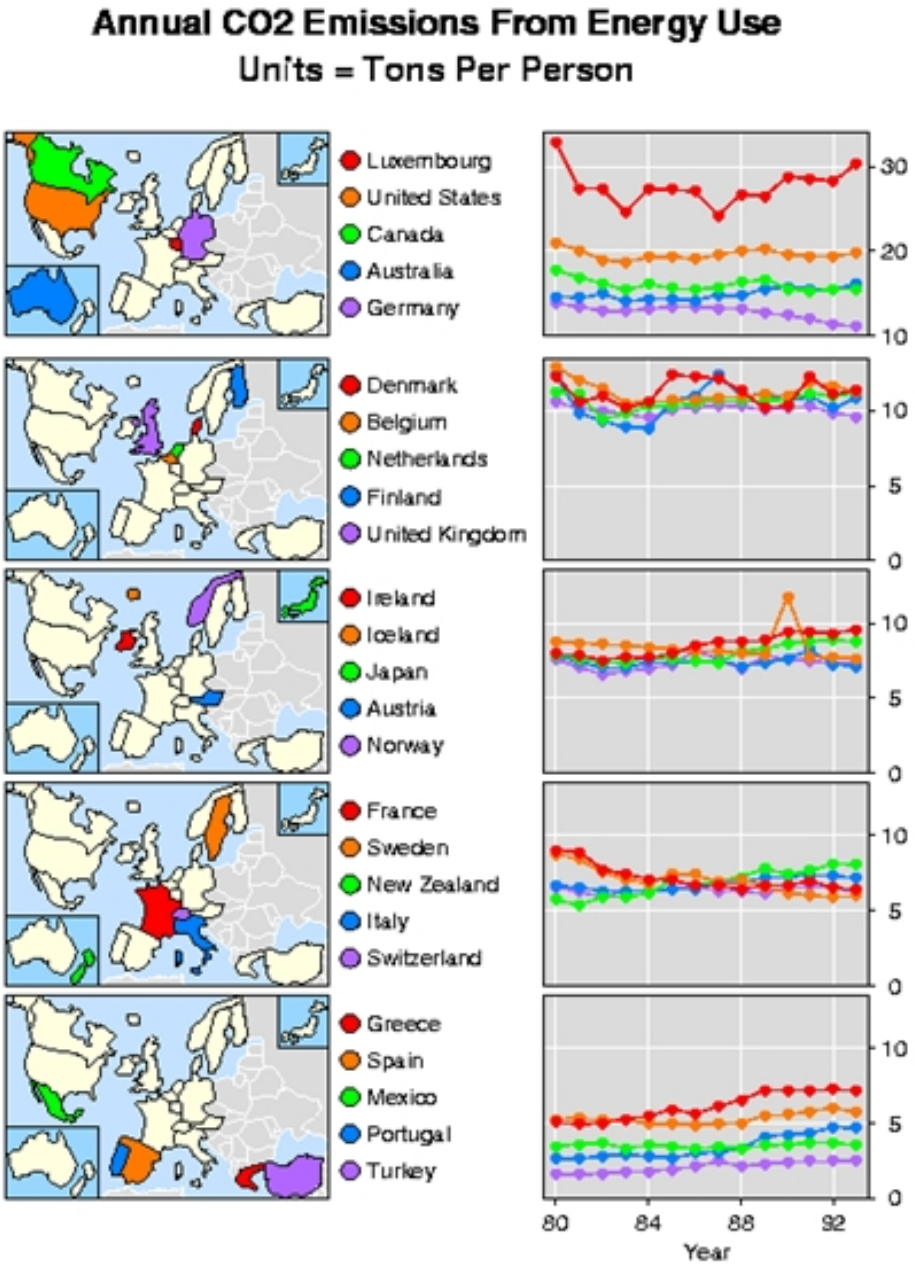


Figure 99: Example from Dan Carr, showing CO2 emissions over time.

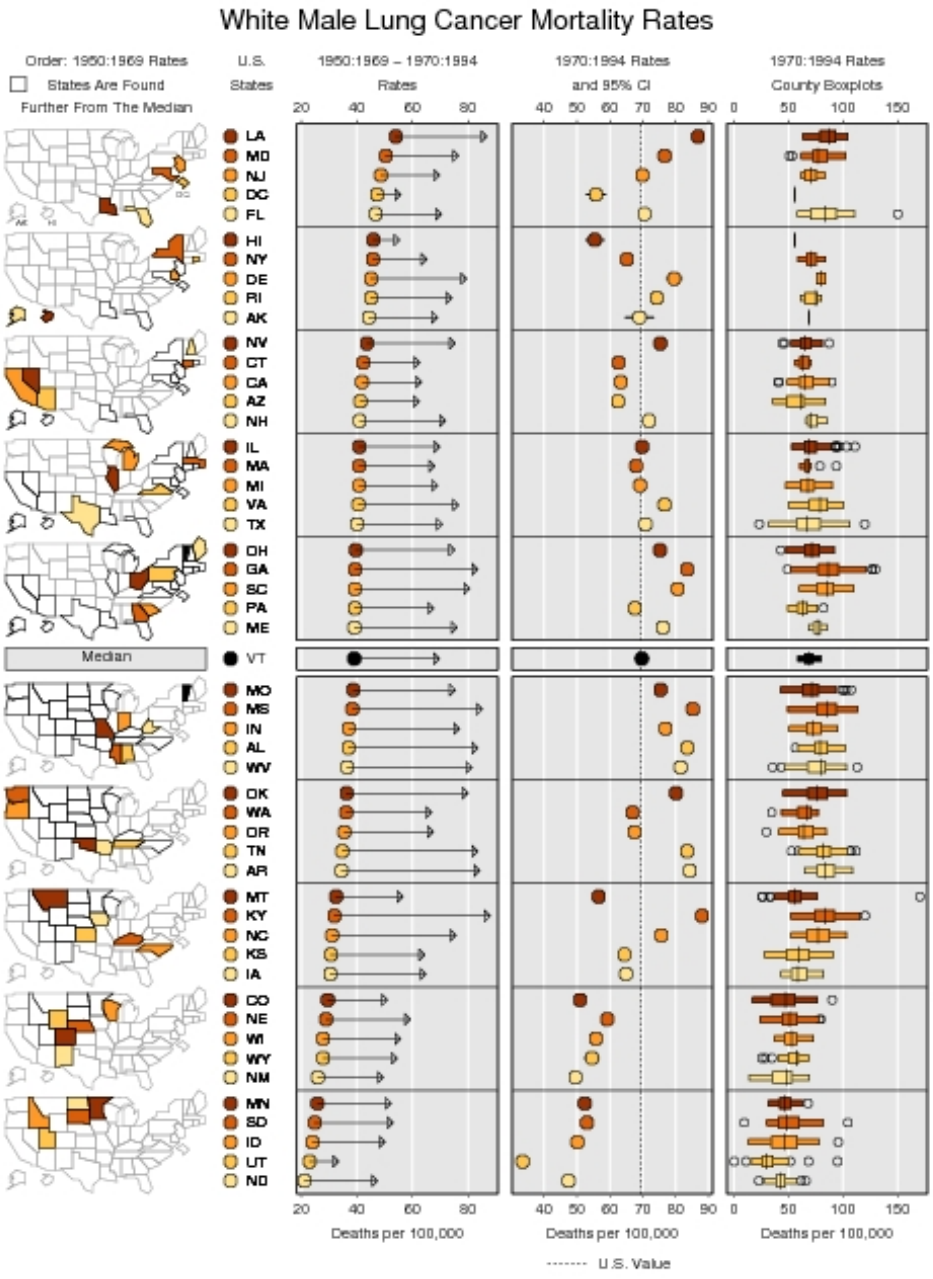


Figure 100: Symanzik & Carr (2008), p. 286, Figure 1.6: LM plots, based on data from the NCI Web page, showing summary values for the years 1950 to 1969 and for the years 1970 to 1994 in the left data panel, rates and 95% confidence intervals in the middle data panel, and boxplots for each of the counties of each state in the right data panel.

9.3.4 Web–based Applications of LM Plots

Symanzik & Carr (2008), pp. 276–282, state:

> "Over the last decade, U.S. Federal Agencies and other institutions increasingly focused attention on distributing large amounts of geographically referenced statistical data, either in print or through the Web. The Web–based distribution of data is aimed at replacing printed tabular displays and at providing access to current data quickly. Several approaches have been developed that provide a user–friendly Web–based interface to tabular and graphical displays of Federal data. The user can interactively and dynamically query and sort the data, compare different geographic regions, and look at the data at different spatial resolutions, e.g., at the state or the county level. Carr & Olsen (1996) provide examples on the visual appearance of patterns in data when properly sorted.
>
> The direction of LM plots development shifted from static LM plots towards interactive micromap displays for the Web. Work done for the EPA CEP Web site (Symanzik et al. 2000) was the first in this regard. This project was soon followed by Web–based examples of micromaps produced by the USDA–NASS such as in Fig. 94.
>
> The Digital Government (dg.o) initiative (`http://www.diggov.org`)[‡] is a major research initiative funded by the National Science Foundation (NSF) and several Federal Agencies such as the EPA, the USDA–NASS, the U.S. Census Bureau, the NCI, the U.S. Bureau of Labor Statistics (BLS), etc. This initiative addresses multiple aspects related to Federal data such as visualization, access, disclosure, security, etc. One of the proposals funded under dg.o was the Digital Government Quality Graphics (DGQG) project that included the development of LM plots (`http://www.geovista.psu.edu/grants/dg-qg/index.html`).
>
> In the reminder of this section, we look at four main applications of interactive Web–based LM plots, three of them on Federal Web sites. A short overview of interactive micromaps, as well as a micromap of the "Places" data (Boyer & Savageau 1981), can be found in Symanzik (2004). However, additional details are given in this section.

[‡]Archived versions of these Web pages can be found at `http://web.archive.org/web/*/http://www.diggov.org`.

Micromaps at the EPA CEP Web Site

The idea of using micromaps on the Web was first considered for the EPA CEP Web site (previously accessible at `http://www.epa.gov/CumulativeExposure/`).[§] Initially, the EPA wanted to provide fast and convenient Web–based access to its hazardous air pollutant (HAP) data for 1990. In this data set, concentrations of 148 air pollutants were estimated for each of the 60,803 U.S. census tracts in the 48 contiguous U.S. states (Rosenbaum et al. 1999). The EPA Web site was designed to allow the user to easily move through the data set to find information on different air pollutants at different geographic locations and at different levels of geographic resolution (e.g., state, county, census tract) via interactive tables and micromaps. Unfortunately, no part of the interactive CEP Web site was ever published due to concerns that the 1990 data was outdated at the intended release date in 1998. Only a static version of the CEP Web site without tables and micromaps was accessible for several years. More details on the work related to the planned interactive CEP Web site can be found in Symanzik et al. (2000), Symanzik, Axelrad, Carr, Wang, Wong & Woodruff (1999), Symanzik, Carr, Axelrad, Wang, Wong & Woodruff (1999).

Micromaps at the USDA–NASS Web Site

The USDA–NASS Research and Development Division released a Web site (`http://www.nass.usda.gov/research/sumpant.htm`) in September 1999 that uses interactive micromaps to display data from the 1997 Census of Agriculture. The USDA–NASS Web site displays acreage, production, and yield of harvested cropland for corn, soybeans, wheat, hay, and cotton. [...]

Micromaps at the NCI Web Site

The NCI released the State Cancer Profiles Web site in April 2003 that provides interactive access to its cancer data via micromaps. This Web site is Java–based and creates micromaps "on the fly". Wang et al. (2002) and Carr

[§]Archived versions of these Web pages can be found at `http://web.archive.org/web/*/http://www.epa.gov/CumulativeExposure/`.

et al. (2002) provide more details on the design of the NCI Web site that is accessible at `http://www.statecancerprofiles.cancer.gov/micromaps`. [...]

Micromaps at Utah State University

Micromaps and other graphical displays were found to be very useful for the display and analysis of the geographic spread of the West Nile Virus (WNV) and other diseases (Symanzik et al. 2003) across the U.S. For this reason, researchers at Utah State University (USU) obtained the NCI Java micromap code and adapted it for the display of WNV data (Chapala 2005). Similar to the NCI micromap application, a user can now select among WNV infection rates and counts, and death rates and counts, starting with the WNV data for the U.S. for the 2002 season. A drill–down into U.S. counties is possible given that data at the county–level is available. New features of the USU Web site include the plotting of the data for two years side–by–side in one panel and additional sorting criteria such as sorting from the highest increase over no change to highest decrease in two selected years. The USU WNV micromap Web site can be accessed at `http://webcat.gis.usu.edu:8080/index.html`. [...]"

9.4 Micromaps in Java

The National Cancer Institute (NCI) states at `http://gis.cancer.gov/tools/micromaps/`:

> "Linked Micromaps is a graphing program written in Java. It allows users to view multiple variables interactively and compare statistics across regions (states, counties, registries, hospitals) as well as across time. It supports six types of graph:
>
> - bar graphs;
> - box plots;
> - raw data tables;
> - point graphs;
> - point graphs with arrow; and
> - point graphs with confidence intervals.
>
> In order to use Linked Micromaps, you must have Java installed on your PC. Your input files must be in a delimited (such as Comma-Separated Values [CSV]) or fixed–width text format."

9.5 Micromaps in R

Example 1:

Execute this code to create a simplified micromap with one statistical panel, written by Mike Minnotte:

```
library(RColorBrewer)
library(maps)

display.brewer.pal(6, "Set1")
pal <- brewer.pal(6, "Set1")
pal[6] <- "#DDDDDD"

data(state)
murder <- state.x77[,5]
murder.name <- state.name[order(murder, decreasing = T)]
murder.name
murder <- sort(murder, decreasing = T)

layout(matrix(1:36, nrow = 12, byrow = T), widths = c(1, 1, 2),
  heights = c(rep(4, 5), 1, rep(4, 5),3))
layout.show(36)

for (i in 1:10)
{
# compute colors, plot map (column 1)

par(mar=rep(.1,4))
if (i <=5) m.col <- c(rep(pal[6],25),rep(0,25))
else m.col <- c(rep(0,25),rep(pal[6],25))
m.col[(i-1)*5+1:5]<-pal[1:5]
map.m.col<-m.col[match.map("state",murder.name)]
map("state",fill=T,col=map.m.col,border=0)

# plot labels (column 2)

par(mar=rep(.1,4))
```

```
plot(0,0,xlim=c(0,1),ylim=c(0,1),type="n",bty="n",
  xaxt="n",yaxt="n",xlab="",ylab="")
points(rep(.1,5),seq(.9,.1,by=-.2),pch=21,bg=pal[1:5],cex=2)
text(rep(.18,5),seq(.9,.1,by=-.2),murder.name[(i-1)*5+1:5],pos=4,cex=1.5)

# plot dotplot of values (column 3)

par(mar=rep(.1,4))
if (i==10) plot(0,0,xlim=range(murder),ylim=c(0,1),type="n",
yaxt="n",xlab="",ylab="")
else plot(0,0,xlim=range(murder),ylim=c(0,1),type="n",xaxt="n",
yaxt="n",xlab="",ylab="")
abline(h=seq(.9,.1,by=-.2),lty=3,col="grey")
points(murder[(i-1)*5+1:5],seq(.9,.1,by=-.2),pch=21,bg=pal[1:5],cex=2)

# separate states above and below median

if (i==5) {for (j in 1:3){
plot(0,0,xlim=c(0,1),ylim=c(0,1),type="n",bty="n",
  xaxt="n",yaxt="n",xlab="",ylab="")
abline(h=.5,lwd=3,col=pal[6])
}}

# Plot through remaining (empty) cells

if (i==10) for (j in 1:3)
plot(0,0,xlim=c(0,1),ylim=c(0,1),type="n",bty="n",
  xaxt="n",yaxt="n",xlab="",ylab="")
}

# Label for dot plots

text(0.5, 0.25, "Murders per 100K Population", cex = 1.5)
```

Example 2:

More sophisticated R code for micromaps, created by Dan Carr. The original version of this R code can be found at `http://classweb.gmu.edu/dcarr/eda/schedule.html` (Week 6). The R code posted here has been modified to run from our course Web page.

```
# Load R Functions

load(url("http://www.math.usu.edu/~symanzik/teaching/2009_stat6560/RDataAndScripts/Ch9_panelLayout.Rdata"))


# Load Data

stateUnemploy95 =
  read.csv(url("http://www.math.usu.edu/~symanzik/teaching/2009_stat6560/RDataAndScripts/Ch9_stateUnemployment95.csv"),
  row.names = 1, header = T)

stateNamesFips =
  read.csv(url("http://www.math.usu.edu/~symanzik/teaching/2009_stat6560/RDataAndScripts/Ch9_stateNamesFips.csv"),
  row.names = 1, header = T)

stateVBorders =
  read.csv(url("http://www.math.usu.edu/~symanzik/teaching/2009_stat6560/RDataAndScripts/Ch9_stateVisibilityBorders.csv"),
  row.names = NULL, header = T)

nationVBorders =
  read.csv(url("http://www.math.usu.edu/~symanzik/teaching/2009_stat6560/RDataAndScripts/Ch9_nationVisibilityBorders.csv"),
  blank.lines.skip = F, row.names = NULL, header = T)


# Create pdf (or jpg) Output

pdf("Ch9_micromap.pdf", width = 7.5, height = 10)
#jpeg("Ch9_micromap.jpg", width = 7.5, height = 10, units = "in", res = 72)

source(url("http://www.math.usu.edu/~symanzik/teaching/2009_stat6560/RDataAndScripts/Ch9_linked_micromaps.R"))

dev.off()
```

9.6 Further Reading

Additional sources for choropleth maps and micromaps are:

- Work related to Pickle et al. (1996): Pickle & Herrmann (1999), Wainer (2008), and Pickle (2008)

10 Interactive and Dynamic Graphics

Symanzik (2004), p. 295, states:

> "Interactive and dynamic statistical graphics enable data analysts in all fields to carry out visual investigations leading to insights into relationships in complex data. Interactive and dynamic statistical graphics involve methods for viewing data in the form of point clouds or modeled surfaces. Higher–dimensional data can be projected into one–, two– or three–dimensional planes in a set of multiple views or as a continuous sequence of views which constitutes motion through the higher–dimensional space containing the data.
>
> Strictly, there is some difference between interactive graphics and dynamic graphics. When speaking of interactive graphics only, we usually mean that a user actively interacts with, i.e., manipulates, the visible graphics by input devices such as keyboard, mouse, or others and makes changes based on the visible result. When speaking of dynamic graphics only, we usually mean that the visible graphics change on the computer screen without further user interaction. An example for interactive graphics might be the selection of interval lengths and starting points when trying to construct an optimal histogram while looking at previous histograms. An example for dynamic graphics might be an indefinitely long grand tour with no user interaction. Typically, interactive graphics and dynamic graphics are closely related and we will not make any further distinction among the two here and just speak of interactive and dynamic statistical graphics.
>
> Two terms closely related to interactive and dynamic statistical graphics are exploratory data analysis (EDA) and visual data mining (VDM).
>
> EDA, as defined by Tukey (1977), "*is detective work — numerical detective work — or counting detective work — or graphical detective work.*" Modern techniques and software in EDA, based on interactive and dynamic statistical graphics, are a continuation of Tukey's idea to use graphics to find structure, general concepts, unexpected behavior, etc. in data sets by looking at the data. To cite Tukey (1977) again, "*today, exploratory and confirmatory can — and should — proceed side by side.*" Interactive and dynamic statistical graphics should not replace common analytic and inferential statistical methods — they should rather extend these classical methods of data analysis."

Due to time constraints, we are not able to go deeper into interactive and dynamic graphics this semester.

10.1 Further Reading

Additional sources for interactive and dynamic graphics are:

- GGobi: Swayne et al. (2003)
- Mondrian: Theus (2002, 2003), Theus & Urbanek (2009)
- iPlots: Theus & Urbanek (2004)

11 Graphics Galleries and Sources on the Web

Graphics Galleries:

- Michael Friendly' Statistics and Statistical Graphics Resources: `http://www.math.yorku.ca/SCS/StatResource.html`
- Romain François ' R Graph Gallery: `http://addictedtor.free.fr/graphiques/`
- Paul Murrell's R Graphics Web page, accompanying Murrell (2006): `http://www.stat.auckland.ac.nz/~paul/RGraphics/rgraphics.html`

Social Data Analysis Web Sites:

- NameVoyager - Baby Name Wizard of Most Popular Baby Names (Martin and Laura Wattenberg): `http://www.babynamewizard.com/voyager`
- Many Eyes (IBM): `http://many-eyes.com`
- Swivel: `http://www.swivel.com/`
- StatCrunch (Webster West): `http://statcrunch.com/`
- Talks: Hans Rosling: Debunking Third–World Myths with the Best Stats You've Ever Seen (TED.com — featured in Time, April 27, 2009, p. 44): `http://www.ted.com/index.php/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen.html`; see `http://www.gapminder.org/` for more details and examples

Statistical Graphics Section:

- Statistical Computing and Statistical Graphics Sections of the American Statistical Association (ASA) Home: `http://stat-computing.org/`
- Statistical Graphics Video Library: `http://stat-graphics.org/movies/`

Appendix

Homework Assignments

Stat 6560 Graphical Methods	Spring Semester 2009

Homework Assignment 1 (1/12/2009)

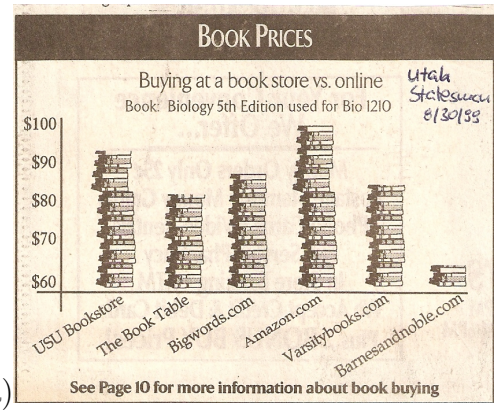
25 (+ 5 EC + $n \times 3$ EC) Points — Due 1/23/2009 (by 1pm)

(i) (10 Points) 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 ...").

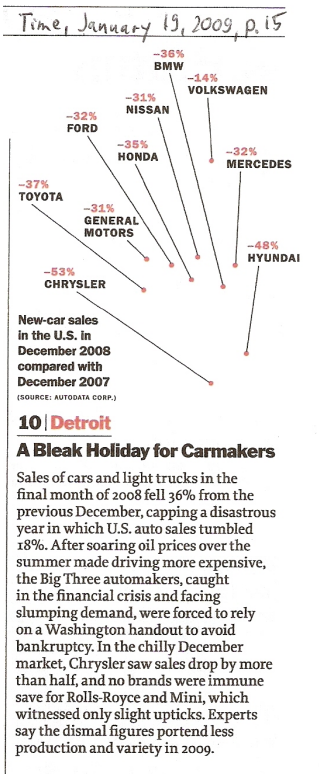
(∗) For each of these figures, explain which rule(s) (how to construct a bad graphic) from our lecture notes the graphics designer has followed, i.e., list the rule(s) and explain why it has been followed. Demonstrate how these poor graphics might be improved. Using the data from the graphic (or your 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) (10 Points) Repeat (∗) for the two recent graphics included below.

(a)

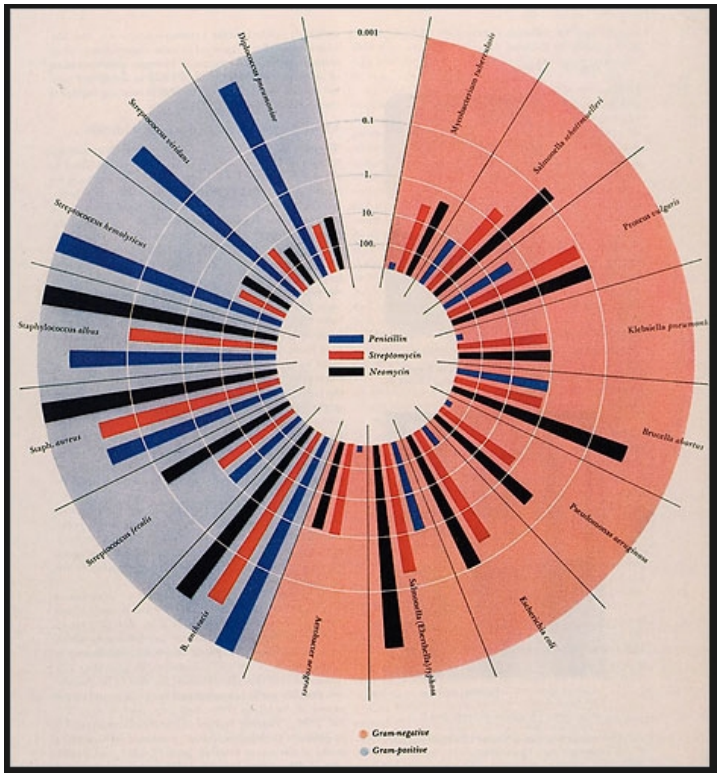


(b)



(iii) (5 Points) 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 specifically on graphics and charts, but from an original (preferably recent) source. Journal articles, newspapers, magazines, and scholarly books are all appropriate sources. Repeat (∗) for your graphic. Turn in a scan of the original together with precise documentation (source and page, URL, etc.) where you found that figure.

(iv) (Individual EC: 5 Points, due 2/20/2009, 1pm) Construct a graphic for the *Chance* graphics contest, as described in *Chance* 21(4), p. 62, from Fall 2008. Also see `http://www.public.iastate.edu/~larsen/` and `http://www.public.iastate.edu/~larsen/graphic%20contest/graphics%20contest.pdf`. The data (in csv format) are available from these Web pages. If you want to participate at the competition, their deadline is 1/15/2009! 2/20/2009 is our internal deadline (just for this course). I could not find the original 1951 figure in a book or article published by Will Burtin, but the NY Times republished the figure as part of a book review, dedicated to his life and work.



Check here `http://www.nytimes.com/2008/06/01/books/review/Heller-t.html?_r=3&oref=slogin` for the book review and the source of the figure above.

(v) (Individual EC: 3 Points each, up to 5 times) For the duration of the course (until 4/17/2009), part (iii) may be repeated up to five times for extra credit. For each example of a poor published graphic which is turned in and handled as in (∗), up to 3 points will be added to your total score.

NOTE: EC questions usually are individual questions. 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 and all other students in class, indicating about the following: *I found a poor graphic on page 1 of the Salt Lake Tribune on Wed 1/7/2009, showing the average tax increase and decrease per person in the case of changes to the state's cigarette tax.* From the time you send this e–mail, you have 7 days to turn in your full answer. If you send an e–mail but do not turn in your answer within 7 days, 2 points will be subtracted from your final score. This is to prevent that someone grabs all easily accessible poor graphics (from sources like CNN, Time magazine, the Utah Statesman, etc.) and prevents others to work on these graphics. Also, at most one entry per person per week.

General Submission Rules: (for homework, EC, and individual reports)

All your submissions this semester must be typeset in LaTeX. In fact, your submissions should translate via `pdflatex`. Figures (from scans or from graphical software) must accompany your 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. LaTeX warnings are OK, but 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. There is no need to submit the final pdf file as I will retranslate all documents on my side.

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

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

lastname_firstname_EC_hwJ_qL_main.tex
lastname_firstname_EC_hwJ_qL_figK.jpg

lastname_firstname_projectN_main.tex
lastname_firstname_projectN_figK.jpg
```

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. Your `groupI_hwJ_main.tex` (and your `lastname_firstname_EC_hwJ_qL_main.tex` and your `lastname_firstname_projectN_main.tex`) need to translate correctly in my `lect_main.tex` environment. All necessary templates will be provided later this week.

Stat 6560 Graphical Methods — Spring Semester 2009

Homework Assignment 2 (2/20/2009)

25 Points — Due 3/4/2009 (by 1pm)

(i) (10 Points) Similar to Section 2.2 in our Lecture Notes, provide a brief (at most one page each) summary of a historical contributor to statistical graphics and of a currently active researcher in statistical graphics. Include a painting or a photo, an important plot, and summarize the main contributions of this person. For a current researcher, you should include a link to this person's personal Web page.

Use Michael Friendly's Milestones Web page at `http://www.math.yorku.ca/SCS/Gallery/milestone/milestone.pdf` to identify suitable candidates for this summary.

Each homework group should deal with different persons than the other groups. You have to coordinate by 2/27/2009 at the latest which group is going to work with which historical and current persons. If two (or more) groups want to provide a summary of the same person, flip a coin!

(ii) (10 Points) Revisit our R code from Section 3.1 from our Lecture Notes at `http://www.math.usu.edu/~symanzik/teaching/2009_stat6560/RDataAndScripts/Ch3_RGBColors.R`. Choose 10 color schemes (of 8 colors each) as follows:

- 2 schemes related to the pure R, G, and B schemes used in class (you only need 8 colors; we used 10 colors in class)
- 2 sequential schemes from RcolorBrewer
- 2 diverging schemes from RcolorBrewer
- 2 sequential schemes from colorspace
- 2 diverging schemes from colorspace

Arrange your 10 plots of the same data, but using these different color schemes, in a 5 rows × 2 column layout on a single page. Clearly label which color scheme has been used for each plot, similar to the ColorBrewer names used in Section 3.4.1 of our Lecture Notes. Write the output into a pdf file via the `pdf` command in R.

Now, consider the following 4 viewing options:

- print the pdf on a color printer
- print the pdf on a black and white printer (or copy the color printout on a black and white copy machine)
- look at the pdf on a PC
- look at the pdf on a laptop

For each of these four different viewing options, individually rank the color visibility. Which plot best shows 8 different colors (1), which is second best (2), ..., which is worst (10). Ties are not allowed.

This ranking should be done by all group members under the same conditions, e.g., at the same computer under the same lighting conditions. Ideally, all group members should simultaneously look at the 10 plots of a particular viewing option and than rank the 10 plots independently.

First tabulate your individual rankings for each of the 10 plots for each of the 4 different viewing options for each of the 2 (or 3) students. Then calculate average ranks for each viewing option and finally calculate the overall average rank over all 4 viewing options. Include these tables as part of your homework submission.

Which color scheme is the best under the 4 viewing options? And which is the best across all 4 viewing options? Is your overall winner also a winner under 1 (or more) of the 4 viewing options, or is the overall winner always highly ranked, but not necessarily an individual winner?

(iii) (5 Points) As this is a Graphics class, also provide a graphical representation of your rankings. This figure should contain rankings for the 10 color scheme choices $\times$ the 4 viewing options $\times$ the 2 (or 3) students. Also display the average ranks for each viewing option and the overall average rank.

General Submission Rules: (for homework, EC, and individual reports)

Please follow the general submission rules from Homework 1. In addition, please follow these instructions:

- Make sure to include all graphical results into your pdf submission. Link to the R files, assuming they are in the same directory as the main tex file.
- Include a title page, including course name, student names, number of homework, and the submission date.
- Start answers to each question on a new page, clearly labeling which question is being answered.
- Define a toggle (say `tasksbystudent`) that allows you to switch on/off the following information:

  On the very last page of your submission, provide a table that summarizes for each question how much each group member has contributed to each of the main tasks of that question, e.g., Question x: R coding (Student A: 80%, Student B: 20%), LaTeX typesetting (A: 10%, B: 90%), Other tasks, such as selection of color schemes (A: 50%, B: 50%).

  It is totally up to you how to split tasks, even up to the extreme case that one student does 100% of all tasks for one question and, thus, another student does 100% of all tasks for the next question. However, each student should do more than 50% of the work for each of the main tasks (in particular R coding and LaTeX typesetting) at some point. So, a submission where one student does 100% (or just 51%) of all R coding for all questions on an assignment while another student does 100% (or just 51%) of all LaTeX typesetting for all questions is *not* acceptable.

  Thus, decide in advance who will take the lead for each of the main tasks on a particular assignment sheet. If I nevertheless see such an uneven distribution of tasks, I will ask the trailing student(s) to work on some additional questions to become more familiar with a particular main task of the homework assignment.

Stat 6560 Graphical Methods Spring Semester 2009

Homework Assignment 3 (3/20/2009)

40 Points — Due 4/6/2009 (by 1pm)

(i) (5 Points) Online documentation and help files for statistical software are often incomplete. The R help page for mosaicplot is not clear at all on how the standardized residuals are calculated. The only information provided is: "Extended mosaic displays visualize standardized residuals of a loglinear model for the table by color and outline of the mosaic's tiles. (Standardized residuals are often referred to a standard normal distribution.) Negative residuals are drawn in shaded of red and with broken outlines; positive ones are drawn in blue with solid outlines."

Search for additional literature on mosaic plots, e.g., starting with the references from the R help page for mosaicplot, and try to figure out how *exactly* the standardized residuals are being calculated.

Verify this for the HairEyeColor data set, i.e., create a table that lists all possible combinations for hair color, eye color, and gender and then indicate the exact numerical value for each standardized residual. Which color coding (and outline) should be used for the standardized residual in each case? And which color coding (and outline) is used in the actual R plot:

```
mosaicplot(HairEyeColor, shade = T)
```

(ii) (5 Points) Similar to the previous question, the standardization for fourfoldplots is not documented well in the R help files. The only information provided is: "std: a character string specifying how to standardize the table. Must be one of "margins", "ind.max", or "all.max", and can be abbreviated by the initial letter. If set to "margins", each 2 by 2 table is standardized to equate the margins specified by margin while preserving the odds ratio. If "ind.max" or "all.max", the tables are either individually or simultaneously standardized to a maximal cell frequency of 1."

Again, find appropriate literature and manually determine how the cell radius is being calculated for each of the three std options. In particular, do your calculations for the UCBAd subset of the UCBAdmissions data set. Tabulate the

calculated radius, a measurement of the actual radius in the plot, and list the ratio (calculated radius / measured radius) for each cell for each of the three plots:

```
UCBAd = margin.table(UCBAdmissions, 1:2)
fourfoldplot(UCBAd, std = "m") # m(argins) - the default
fourfoldplot(UCBAd, std = "i") # i(nd.max)
fourfoldplot(UCBAd, std = "a") # a(ll.max)
```

(iii) (25 Points) Read Appendix B: "Washing — What Makes the Difference" in Theus and Urbanek (2009), pp. 165–172. The data can be obtained from `http://www.interactivegraphics.org/Datasets.html`.

(a) (5 Points) Recreate the 13 figures from pp. 168–169 in Mondrian and arrange them in a similar layout as in the book. You may have to resort the columns or rows of your plots or otherwise modify the layout in Mondrian to reproduce exactly the figures shown in the book. The default layout may not match what is shown in the book. Finally, create some screenshots (2 or 3 may be necessary) of your Mondrian displays (similar to Figure 85 in our Lecture Notes that recreates the figures on p. 186 in Theus and Urbanek (2009)).

(b) (5 Points) Switch to R (p. 171 indicates how the data set can be recreated in R) and recreate these two mosaicplots in R, using exactly the same layout as in the Mondrian figures in the book: (a) Water Softness × Temperature, and (b) Temperature × Preference × M–User.

(c) (10 Points) Answer Exercises 1 and 2 on p. 172 in Theus and Urbanek (2009).

(d) (5 Points) Provide a 1 to 2 page summary and discuss the graphical and numerical results from your previous answers given in this question.

(iv) (5 Points) Work with the Old Faithful data set (faithful) that is available in R. Create at least four different bivariate plots that show eruption time (length in minutes; x–axis) and waiting time to next eruption (length in minutes; y–axis). What do we learn from these plots? Summarize your results. Then, compare your results with those shown in Figure 1 in Zeileis, Hornik, and Murrell (2008) [`http://statmath.wu-wien.ac.at/~zeileis/papers/Zeileis+Hornik+Murrell-2008.pdf`]. So, is it sufficient to make a statement of the form "Our figure shows a scatterplot of the Old Faithful data set."? If not, what else should always be indicated?

Please follow the general submission rules as stated on the previous homework assignments.

Stat 6560 Graphical Methods Spring Semester 2009

Homework Assignment 4 (4/20/2009)

25 Points — Due 4/27/2009 (by 5pm)

(i) (25 Points) This homework assignment is related to Bill Morphet's guest lecture in our "Graphical Methods" course, accessible at `http://www.math.usu.edu/~symanzik/teaching/2009_stat6560/Lecture_04_06_2009.pdf`. You will have to answer several survey questions related to arrow plots and circular dataimages. The goal of this survey is to determine which tasks can be solved better (and faster) via arrow plots and which tasks can be solved better (and faster) via circular dataimages. Conducting such a survey was one of the reviewer requests for a journal paper on circular dataimages, based on one of the chapters of Bill's dissertation. The actual survey will be distributed into your mailbox later this afternoon.

Please work on this homework assignment individually. In fact, each survey is different, but you may get some hints in case you look at any of the other surveys. Please do not discuss your survey with anybody else but me. As a reference, you can always refer back to Bill's guest lecture notes at `http://www.math.usu.edu/~symanzik/teaching/2009_stat6560/Lecture_04_06_2009.pdf` or to a printout of these notes.

This homework will **not** be graded on correctness, but on **completeness** only. When you answer all questions, you will be awarded the full 25 points. **This includes the recording of starting times and ending times where asked.** Once you have started a question, finish this question as quickly as possible, in particular, please do not walk away from your desk (and do not make any phone calls or get otherwise distracted) once you have started with a particular question. Please do not return to a previously answered question. Please answer all questions sequentially. Some of the questions are open–ended, so you should write down all your thoughts.

For homework grading purposes, I will link the number on the survey sheet with a student name. Bill, who will do the actual analysis of the surveys, will not obtain any of the names of the participating students.

The survey only contains one personal question related to color vision impairment. You are **not** required to answer this question. In fact, if you do not answer this question, you can still obtain the full points when answering the remaining questions on the survey.

The population of interest for this survey are individuals with an advanced knowledge on statistical graphics, such as obtained in a "Graphical Methods" course. Given our small sample size of only 7 students, it would be highly appreciated if everyone agrees to participate. However, if you prefer that your survey will not be passed on to Bill for a detailed statistical analysis, please indicate **"Do not pass on to Bill."** on the front of your survey and I will only use your answers for grading purposes.

Please place your survey answers into my mailbox by Mo 4/27/09, 5pm.

Thank you for participating!

Jürgen & Bill

Project Descriptions

Stat 6560 Graphical Methods Spring Semester 2009

Project 1.1 (1/14/2009)

John Snow and the Cholera Epidemic in London, 1854

50 Points — In–class Presentation 1/28/2009

(i) Read Snow (1936), pp. 36–55, Tufte (1997), pp. 27–37, Wainer (1997), pp. 60–62, and Carvalho et al. (2004). You can use additional references (books, articles, Web pages), but you have to cite these in your summary and add the printed references (books and articles) to the bib file. For example, `http://matrix.msu.edu/~johnsnow/` is a Web site dedicated to the life and work of John Snow. Many other presentations on this topic exist, e.g., Roger Bivand's ppt file at `http://www.bias-project.org.uk/ASDARcourse/unit0_slides.pdf`. The cholera data and some GIS programs can be found at `http://www.ncgia.ucsb.edu/pubs/snow/snow.html`.

(ii) (25 Points) Provide a five (four to six are OK) page summary that contains some background on the medical knowledge in the middle of the 19th century, in particular how cholera was assumed to spread. Explain the contribution of John Snow. Why was this so outstanding and is now considered as the foundation of modern epidemiology?

Choose one or two of the graphics you present and obtain the underlying data. You may find the data directly in R, in another software package, somewhere on the Web, in the printed references, or you may have to estimate the data as good as possible from the printed references. Then, recreate these figures in R. In case R code already exists somewhere to recreate these figures, you can obviously start with that code. But, you must fully understand what each line of the code (i.e., each command and each function) you found is doing.

Your report should be written in LaTeX and should translate in the general LaTeX framework provided earlier in the semester. All figures need to be included in electronic format. Don't forget to list the references you have used. Do not include your R code in your five pages, but rather link to it as done earlier in the semester. Follow our general submission rules. Your written report is due on 1/26/2009 before class.

(iii) (25 Points) Give a 20 min presentation of your project in class on 1/28/2009. You should provide an interesting presentation that is based on your written report. Introduce the general background and carefully discuss the figures from your report. Demonstrate how one or two of the graphics you have presented can be recreated using R. Make sure you can explain the general outline of your R code, but also be prepared to explain what individual R commands or functions you use to create your figures are for.

Stat 6560 Graphical Methods Spring Semester 2009

Project 1.2 (1/14/2009)

The Challenger Disaster in 1986: How Graphics played a Deadly Role

50 Points — In–class Presentation 1/30/2009

(i) Read Tufte (1997), pp. 27 & 38–53, and Wainer (1997), pp. 51–53. You can use additional references (books, articles, Web pages), but you have to cite these in your summary and add the printed references (books and articles) to the bib file. You may want to start your in–class presentation with a brief (2 to 3 min) news video from `http://www.youtube.com/` that shows the launch and explosion of the Challenger.

(ii) (25 Points) Provide a five (four to six are OK) page summary that contains some background on the American space shuttle program in the 1980ies, the discussions before the Challenger launch, the misleading graphics that were used to justify its launch, and the correct figure that shows why Challenger never should have been launched that day.

Choose one or two of the graphics you present and obtain the underlying data. You may find the data directly in R, in another software package, somewhere on the Web, in the printed references, or you may have to estimate the data as good as possible from the printed references. Then, recreate these figures in R. In case R code already exists somewhere to recreate these figures, you can obviously start with that code. But, you must fully understand what each line of the code (i.e., each command and each function) you found is doing.

Your report should be written in LaTeX and should translate in the general LaTeX framework provided earlier in the semester. All figures need to be included in electronic format. Don't forget to list the references you have used. Do not include your R code in your five pages, but rather link to it as done earlier in the semester. Follow our general submission rules. Your written report is due on 1/28/2009 before class.

(iii) (25 Points) Give a 20 min presentation of your project in class on 1/30/2009. You should provide an interesting presentation that is based on your written

report. Introduce the general background and carefully discuss the figures from your report. Demonstrate how one or two of the graphics you have presented can be recreated using R. Make sure you can explain the general outline of your R code, but also be prepared to explain what individual R commands or functions you use to create your figures are for.

Stat 6560 Graphical Methods Spring Semester 2009

Project 1.3 (1/16/2009)

William Playfair: A Graphical Pioneer of the 18th Century

50 Points — In–class Presentation 2/6/2009

(i) William Playfair is called *"the father of modern graphical display"*, *"the great pioneer of statistical graphics"*, and *"the founder of graphical methods of statistics"*. There exist numerous publications dedicated to this father–pioneer–founder.

Read Wainer (2005), pp. 5–38, Spence (2006), and extracts from Playfair (2005). You can use additional references (books, articles, Web pages), but you have to cite these in your summary and add the printed references (books and articles) to the bib file. Additional details such as pdf files, color versions of Playfair's graphics, and many additional references can be found at Ian Spence's Web page at `http://www.psych.utoronto.ca/users/spence/Research_WP.html` and at the Wikipedia page on Playfair at `http://en.wikipedia.org/wiki/William_Playfair`.

(ii) (25 Points) Provide a five (four to six are OK) page summary that contains some background on Playfair's life and work. Start with citations how others rate Playfair, including the ones quoted above. Then put Playfair in context, i.e., what was done before him? What were his main graphical innovations? Select a few examples that show the breadth of his graphical work.

But, even fathers and pioneers are not always perfect. We have seen one problem with one of Playfair's graphics in class already. In Playfair (2005), pp. 18–23, a few flaws in Playfair's graphics are described. The color versions of these figures (actually omitted in the print issue) can be found at `http://www.psych.utoronto.ca/users/spence/Errata%20Playfair%20Facsimile%20Edition.pdf`. Mention one or two of these flaws, but do not overemphasize on those as his graphical innovations considerably outweigh the few mistakes he made.

Choose one or two of the graphics you present and obtain the underlying data. You may find the data directly in R, in another software package, somewhere on the Web, in the printed references, or you may have to estimate the data as good

as possible from the printed references. Then, recreate these figures in R. In case R code already exists somewhere to recreate these figures, you can obviously start with that code. But, you must fully understand what each line of the code (i.e., each command and each function) you found is doing.

Your report should be written in LaTeX and should translate in the general LaTeX framework provided earlier in the semester. All figures need to be included in electronic format. Don't forget to list the references you have used. Do not include your R code in your five pages, but rather link to it as done earlier in the semester. Follow our general submission rules. Your written report is due on 2/4/2009 before class.

(iii) (25 Points) Give a 20 min presentation of your project in class on 2/6/2009. You should provide an interesting presentation that is based on your written report. Introduce the general background and carefully discuss the figures from your report. Demonstrate how one or two of the graphics you have presented can be recreated using R. Make sure you can explain the general outline of your R code, but also be prepared to explain what individual R commands or functions you use to create your figures are for.

Stat 6560 Graphical Methods Spring Semester 2009

Project 1.4 (1/28/2009)

Charles Joseph Minard and *the Best Statistical Graphic Ever Drawn*

50 Points — In–class Presentation 2/13/2009

(i) Charles Joseph Minard's plot of Napoleon's Russian Campaign is often called *"the best statistical graphic ever drawn"* (Tufte 1983, p.40).

Read Tufte (1983), pp. 40–41 and more, Wainer (1997), pp. 63–65, Robinson (1967), Hankins (1999), and Friendly (2000*a*) (a pdf of the newsletter is available at `http://stat-computing.org/newsletter/v111.pdf`). You can use additional references (books, articles, Web pages), but you have to cite these in your summary and add the printed references (books and articles) to the bib file. Additional details such as pdf files, color versions of Minard's maps, and many additional references can be found at Michael Friendly's Web pages at `http://www.math.yorku.ca/SCS/Gallery/re-minard.html` and `http://www.math.yorku.ca/SCS/Gallery/minbib/index.htm`, at Edward Tufte's Minard Web page at `http://www.edwardtufte.com/tufte/newet`, and at the Wikipedia page on Minard at `http://en.wikipedia.org/wiki/Charles_Joseph_Minard`. R code and data provided by Hadley Wickham for Napoleon's Russian Campaign plot are available from Michael Friendly's Web page.

(ii) (25 Points) Provide a five (four to six are OK) page summary that contains some background on Minard's life and work. Start with citations how others rate Minard, including the one quoted above. Then put Minard in context, i.e., what was done before him? What were his main graphical innovations? Select a few examples that show the breadth of his graphical work.

Then focus on two of Minard's graphs: Napoleon's Russian Campaign plot and a graph of your choice (the commercial movement of merchandise on the Canal du Centre in 1844 might be easy to reconstruct). Obtain the underlying data. You may find the data directly in R, in another software package, somewhere on the Web (such as R code and data provided by Hadley Wickham for Napoleon's Russian Campaign plot), in the printed references, or you may have to estimate the data as good as possible from the printed references. Then, recreate these

figures in R. In case R code already exists somewhere to recreate these figures, you can obviously start with that code. But, you must fully understand what each line of the code (i.e., each command and each function) you found is doing.

Your report should be written in LaTeX and should translate in the general LaTeX framework provided earlier in the semester. All figures need to be included in electronic format. Don't forget to list the references you have used. Do not include your R code in your five pages, but rather link to it as done earlier in the semester. Follow our general submission rules. Your written report is due on 2/11/2009 before class.

(iii) (25 Points) Give a 20 min presentation of your project in class on 2/13/2009. You should provide an interesting presentation that is based on your written report. Introduce the general background and carefully discuss the figures from your report. Demonstrate how two of the graphics you have presented can be recreated using R. Make sure you can explain the general outline of your R code, but also be prepared to explain what individual R commands or functions you use to create your figures are for.

Stat 6560 Graphical Methods Spring Semester 2009

Project 1.5 (2/6/2009)

Visual Perception and Change Blindness

50 Points — In–class Presentation 2/20/2009

(i) Read Rensink (2006) as a general overview of your project and its relevance to the design of statistical graphics. Then focus on the main topic of your presentation, i.e., change blindness (and change detection). Starting at Ronald A. Rensink's home page at `http://www.psych.ubc.ca/~rensink/`, see which useful information you can find in his contributions at `http://www.psych.ubc.ca/~rensink/conDescr.html`. Perhaps this paper at `http://www.psych.ubc.ca/~rensink/publications/abs.05.0.html` might serve as a good starting point. However, you should decide yourself which of his publications on change blindness, available at `http://www.psych.ubc.ca/~rensink/publications/index.html`, are worth detailed reading for this project. Most abstracts and many pdfs of his publications are accessible via this Web page. There are also many examples (as movies) available via his Web pages.

You should use some of these additional references (books, articles, Web pages), and you have to cite these in your summary and add the printed references (books and articles) to the bib file.

(ii) (25 Points) Provide a five (four to six are OK) page summary of what you have read. Start with a brief overview why knowledge about visual perception is important for the effective design of statistical graphics, following Rensink (2006). Then move on to the main topic of your project, i.e., change blindness (and change detection), following the references outlined above.

Include a few sketches and figures in your presentation. The main demonstration of change blindness has to be made in R. Learn more about the R package pixmap at `http://cran.r-project.org/web/packages/pixmap/pixmap.pdf`. Install and load the R packages pixmap and RGraphics on your computer. Then access the R code for Figure 3.26 at `http://www.stat.auckland.ac.nz/~paul/RGraphics/chapter3.html` that is accompanying Murrell (2006), pp. 106–107. Make sure that this example is working on your computer.

Now, write some R code that shows one figure, then a black (or white) empty screen, a modified version of this figure, then a black (or white) empty screen again, etc. — and repeat until this is interrupted by a user. For the figures, use some of Rensink's original examples that can be accessed via his movies (or related publications). Finally, create a pure statistical graphic where one version is slightly modified, e.g., one outer class is missing in a histogram, a line is missing in a time series plot of 5 or 6 variables, or the legend, axis label, or main title are missing from a figure. Also alternate these two versions of the statistical figure in your general R code.

Please note that the R package pixmap works with bitmaps (pnm files). However, Windows bmp files do not work with this package. Check the Web page dedicated to PNM files at `http://people.sc.fsu.edu/~burkardt/data/pnm/pnm.html` to learn more about this file type. There are links to different software packages that allow to create PNM files that can be used as input to the R package pixmap. Before you experiment with Rensink's figures and your own figures, you may want to experiment with the given code and some of the sample PNM figures from the URL above to produce something like Figure 101:

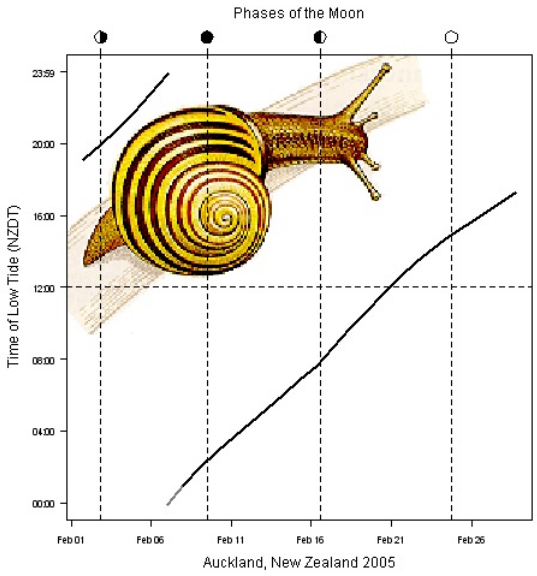


Figure 101: Combination of R code, taken from `http://www.stat.auckland.ac.nz/~paul/RGraphics/custombase-pixmap.R`, with a different PNM figure, taken from `http://people.sc.fsu.edu/~burkardt/data/pnm/snail2.pnm`, on 2/5/2009.

Your report should be written in LaTeX and should translate in the general LaTeX framework provided earlier in the semester. All figures need to be included in electronic format. Don't forget to list the references you have used. Do not include your R code in your five pages, but rather link to it as done earlier in the semester. Follow our general submission rules. Your written report is due on 2/18/2009 before class.

(iii) (25 Points) Give a 20 min presentation of your project in class on 2/20/2009. You should provide an interesting presentation that is based on your written report. Introduce the general background and carefully discuss the figures from your report. As your overall R code will be very short, you should demonstrate several examples of the effect on change blindness, using given examples from the literature and some statistical graphics. Make sure you can explain the general outline of your R code, but also be prepared to explain what individual R commands or functions you use to create your figures are for.

Stat 6560 Graphical Methods Spring Semester 2009

Project 1.6 (2/13/2009)

Andreas Buja: A Modern Pioneer for Interactive Graphics

50 Points — In–class Presentation 2/27/2009

(i) Andreas Buja could be considered as a modern pioneer for interactive graphics.

Start with the interview with Andreas Buja in Symanzik (2008), then look at his contributions to interactive and dynamic statistical software in Symanzik (2004). You can use additional references (books, articles, Web pages), but you have to cite these in your summary and add the printed references (books and articles) to the bib file. Additional details such as his full vita, pdfs of his papers, and videos related to his work can be found at `http://stat-graphics.org/movies/` and at his personal Web page at `http://www-stat.wharton.upenn.edu/~buja/`.

(ii) (25 Points) Provide a five (four to six are OK) page summary that contains some background on Andreas Buja's academic life and work. Based on the interview, how does he see and rate Statistical Graphics? What about teaching Statistical Graphics at a university? Then concentrate on his graphical work, in particular his contributions to the DataViewer, XGobi, and GGobi software family. What are some of the main features of these packages? What is their current status (think of the RGGobi package at `http://cran.r-project.org/web/packages/rggobi/index.html`). Select a few examples (from papers or screenshots from the videos) that show the development of his graphical work.

Choose two of the graphics you present and obtain the underlying data. You may find the data directly in R, XGobi (`http://www.research.att.com/areas/stat/xgobi/`), GGobi (`http://ggobi.org/`), or in another software package, somewhere on the Web, in the printed references, or you may have to estimate the data as good as possible from the printed references. Then, recreate these figures in R (possibly using the RGGobi package). For example, if you choose a figure from a XGobi paper, obtain the data first. Then read in the data into R, and activate GGobi from within R. Use R to change parameters and options for the graphical appearance of the figure in the GGobi window as far as possible. Ultimately, the reproduced figure should resemble the old figure as closely as pos-

sible. This includes, colors, symbols, etc. However, you only have to concentrate on the graphical component of a figure. GGobi and XGobi are different, so the controls of GGobi and any previous package will appear differently. In case R code already exists somewhere to recreate these figures, you can obviously start with that code. But, you must fully understand what each line of the code (i.e., each command and each function) you found is doing.

Your report should be written in LaTeX and should translate in the general LaTeX framework provided earlier in the semester. All figures need to be included in electronic format. Don't forget to list the references you have used. Do not include your R code in your five pages, but rather link to it as done earlier in the semester. Follow our general submission rules. Your written report is due on 2/25/2009 before class.

(iii) (25 Points) Give a 20 min presentation of your project in class on 2/27/2009. You should provide an interesting presentation that is based on your written report. Introduce the general background and carefully discuss the figures from your report. As a motivation, you may want to show (parts of) some of the videos that are related to Andreas Buja's work. Demonstrate how two of the graphics you have presented can be recreated using R. Make sure you can explain the general outline of your R code, but also be prepared to explain what individual R commands or functions you use to create your figures are for.

Stat 6560 Graphical Methods Spring Semester 2009

Project 1.7 (2/20/2009)

Computer Graphics in Teaching Statistics

50 Points — In–class Presentation 3/6/2009

(i) Start with the historical paper by Gentleman (1977), dealing with the suggested use of computer graphics in teaching statistics. Then look at more diverse articles such as Snell & Peterson (1992), Tijms (1992), Krieger & Pinter-Lucke (1992), Gordon & Gordon (1992), and Holcomb & Spalsbury (2005). You can use additional references (books, articles, Web pages), but you have to cite these in your summary and add the printed references (books and articles) to the bib file. Recent conferences dedicated to teaching statistics, such as *OZCOTS 2008 – 6th Australian Conference on Teaching Statistics* at `http://silmaril.math.sci.qut.edu.au/ozcots2008/index.html` might provide many recent publications regarding the use of computer graphics in teaching statistics.

(ii) (25 Points) Provide a five (four to six are OK) page summary that contains an overview on the use of computer graphics in teaching statistics. What are some of the main applications for such graphics, e.g., probability distributions, sampling distributions, etc.? Now, more than 30 years later, which of these early suggestions made by Gentleman (1977) can be found in our everyday classes at Utah State University? Only look at those classes you have attended or taught yourself (from Stat1040 to Stat 6710/20).

Demonstrate two of the concepts suggested in the literature using R. In case R code already exists somewhere for these concepts, you can obviously start with that code. But, you must fully understand what each line of the code (i.e., each command and each function) you found is doing.

Your report should be written in LaTeX and should translate in the general LaTeX framework provided earlier in the semester. All figures need to be included in electronic format. Don't forget to list the references you have used. Do not include your R code in your five pages, but rather link to it as done earlier in the semester. Follow our general submission rules. Your written report is due on 3/4/2009 before class.

(iii) (25 Points) Give a 20 min presentation of your project in class on 3/6/2009. You should provide an interesting presentation that is based on your written report. Introduce the general background and carefully discuss the concepts from your report. Demonstrate how two of the graphical concepts you have presented can be recreated using R. Make sure you can explain the general outline of your R code, but also be prepared to explain what individual R commands or functions you use to demonstrate your concepts are for.

Stat 6560 Graphical Methods Spring Semester 2009

Project 2 (3/20/2009)

Specialized Graphics in R

50 Points — In–class Presentation 4/10/2009–4/22/2009

In Project 2, you will have to present an R package for a particular statistical discipline and focus on the graphics in this package. What are those graphics used for, how can we interpret what we see in these graphics, and what are some design choices for these graphics?

You have to prepare lecture notes related to your R package, similar to what we have done since Chapter 4 in our course Lecture Notes. You should briefly summarize the main graphical commands for this package and introduce the data set you use. Then create a series of R commands that demonstrate the creation and interpretation of the main graphics.

In class, you have to work through your prepared material in 18 to 20 minutes. You will have at most 20 minutes for your in–class preparation as we will schedule presentations of two R packages in the same lecture.

With respect to data, you are **not** allowed to use any data set that currently exists in R. Please check carefully! Instead, you have to use data that originate from a textbook, from a journal paper or a conference paper, or that are available through some general Web site. You are also **not** allowed this time to simply reproduce a previously published figure. Instead, you have to select a new data set and apply the R commands to this new data set. As an example, when you work on regression graphics, the data sets that accompany the two main references in (iii) below are prohibited. Instead, you should find some meaningful data set that is introduced in a different regression textbook (or paper) and then apply the graphical methods introduced in the two references below to the data set you found.

The following choices of R packages and topics are available for Project 2:

(i) R Package *animation*: see `http://animation.yihui.name/start` for more details

(ii) Interactive graphics for R via R Package *iplots*: see Theus and Urbanek (2004) paper at `http://stat-computing.org/newsletter/v151.pdf`, pp. 11–14, for more details

(iii) Regression Graphics (via R), as described in Cook & Weisberg (1999) [`http://www.stat.umn.edu/arc/`] and Cook (1998) [`http://www.stat.umn.edu/RegGraph/`], e.g., from R packages *dr* [see also Weisberg (2002) paper at `http://www.jstatsoft.org/v07/i01`], robust regression work by David Olive [`http://www.math.siu.edu/olive/ol-bookp.htm`], or *lars*

(iv) Graphics for spatial point patterns, e.g., from R packages *splancs*, *spatgraphs*, or *spatstat* and numerous other R packages summarized at `http://cran.nedmirror.nl/web/views/Spatial.html`

(v) Graphics for spatially continuous data (such as variogram cloud plots, variograms, various kriging surfaces, and standard errors), e.g., from R packages *gstat* or *geoR* and numerous other R packages summarized at `http://cran.nedmirror.nl/web/views/Spatial.html`

(vi) Graphics for genetics and biomedical data (such as heatmaps), e.g., from R packages *heatmap.plus* and numerous other R packages summarized at `http://cran.nedmirror.nl/web/views/Genetics.html`

(vii) Graphics for graphs (networks) (such as social networks, calling circles, "friends", "collaborations", etc.), e.g., from R packages *diagram* or *igraph*

(viii) **Abbass:** R Graphics and numerical R output in a Web browser, via *FastRWeb* [`http://www.rforge.net/FastRWeb/`] and interfacing R via the *RinRuby* software [`http://www.jstatsoft.org/v29/i04`]

References

Andrews, D. F. (1972), 'Plots of High–Dimensional Data', *Biometrics* **28**, 125–136.

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.

Blasius, J. & Greenacre, M., eds (1998), *Visualization of Categorical Data*, Academic Press, San Diego, CA.

Boyer, R. & Savageau, D. (1981), *Places Rated Almanac*, Rand McNally, Chicago, IL.

Brewer, C. A. (1997), 'Spectral Schemes: Controversial Color Use on Maps', *Cartography and Geographic Information Systems* **24**(4), 203–220.

Brewer, C. A. (1999), Color Use Guidelines for Data Representation, *in* '1999 Proceedings of the Section on Statistical Graphics', American Statistical Association, Alexandria, VA, pp. 55–60.

Brewer, C. A. (2003), 'A Transition in Improving Maps: The ColorBrewer Example', *Cartography and Geographic Information Science* **30**(2), 159–162.

Brewer, C. A., MacEachren, A. M., Pickle, L. W. & Herrmann, D. J. (1997), 'Mapping Mortality: Evaluating Color Schemes for Choropleth Maps', *Annals of the Association of American Geographers* **87**(3), 411–438.

Brewer, C. A. & Pickle, L. W. (2002), 'Comparison of Methods for Classifying Epidemiological Data on Choropleth Maps in Series', *Annals of the Association of American Geographers* **92**(4), 662–681.

Brillinger, D. R. (2002), 'John W. Tukey: His Life and Professional Contributions', *The Annals of Statistics* **30**(6), 1535–1575.

Carr, D. B. (1994), Converting Tables to Plots, Technical Report 101, Center for Computational Statistics, George Mason University, Fairfax, VA.

Carr, D. B. (2001), 'Designing Linked Micromap Plots for States with Many Counties', *Statistics in Medicine* **20**(9–10), 1331–1339.

Carr, D. B., Chen, J., Bell, B. S., Pickle, L. & Zhang, Y. (2002), Interactive Linked Micromap Plots and Dynamically Conditioned Choropleth Maps, *in* 'dg.o2002 Proceedings', Digital Government Research Center (DGRC). `http://www.dgrc.org/conferences/2002_proceedings.jsp`.

Carr, D. B., Littlefield, R. J., Nicholson, W. L. & Littlefield, J. S. (1987), 'Scatterplot Matrix Techniques for Large N', *Journal of the American Statistical Association* **82**(398), 424–436.

Carr, D. B. & Nusser, S. M. (1995), 'Converting Tables to Plots: A Challenge from Iowa State', *Statistical Computing and Statistical Graphics Newsletter* **6**(3), 11–18.

Carr, D. B. & Olsen, A. R. (1996), 'Simplifying Visual Appearance by Sorting: An Example using 159 AVHRR Classes', *Statistical Computing and Statistical Graphics Newsletter* **7**(1), 10–16.

Carr, D. B., Olsen, A. R., Courbois, J. P., Pierson, S. M. & Carr, D. A. (1998), 'Linked Micromap Plots: Named and Described', *Statistical Computing and Statistical Graphics Newsletter* **9**(1), 24–32.

Carr, D. B., Olsen, A. R., Pierson, S. M. & Courbois, J. P. (2000), 'Using Linked Micromap Plots to Characterize Omernik Ecoregions', *Data Mining and Knowledge Discovery* **4**(1), 43–67.

Carr, D. B., Olsen, A. R. & White, D. (1992), 'Hexagon Mosaic Maps for Displays of Univariate and Bivariate Geographical Data', *Cartography and Geographic Information Systems* **19**(4), 228–236, 271.

Carr, D. B. & Pierson, S. M. (1996), 'Emphasizing Statistical Summaries and Showing Spatial Context with Micromaps', *Statistical Computing and Statistical Graphics Newsletter* **7**(3), 16–23.

Carr, D. B., Wallin, J. F. & Carr, D. A. (2000), 'Two New Templates for Epidemiology Applications: Linked Micromap Plots and Conditioned Choropleth Maps', *Statistics in Medicine* **19**(17–18), 2521–2538.

Carvalho, F. M., Lima, F. & Kriebel, D. (2004), 'RE: On John Snow's Unquestioned Long Division', *American Journal of Epidemiology* **159**(4), 422.

Chambers, J. M. & Hastie, T. J., eds (1993), *Statistical Models in S*, Chapman & Hall, New York, NY.

Chapala, G. K. (2005), 'Development of Rich Features for Web–Based Interactive Micromaps'. Report, Department of Computer Science, Utah State University.

Chernoff, H. (1973), 'The Use of Faces to Represent Points in k–dimensional Space Graphically', *Journal of American Statistical Association* **68**, 361–368.

Cleveland, W. S. (1985), *The Elements of Graphing Data*, Wadsworth, Monterey, CA.

Cleveland, W. S. (1993), *Visualizing Data*, Hobart Press, Summit, NJ.

Cleveland, W. S. (1994), *The Elements of Graphing Data (Revised Edition)*, Hobart Press, Summit, NJ.

Cook, R. D. (1998), *Regression Graphics: Ideas for Studying Regressions through Graphics*, Wiley, New York, NY.

Cook, R. D. & Weisberg, S. (1999), *Applied Regression Including Computing and Graphics*, Wiley, New York, NY.

Crawford, S. L. & Fall, T. C. (1990), Projection Pursuit Techniques for Visualizing High–Dimensional Data Sets, *in* G. M. Nielson, B. Shrivers & L. J. Rosenblum, eds, 'Proceedings of Visualization in Scientific Computing, Los Alamitos, CA', IEEE Computer Society Press, pp. 94–108.

Dent, B. D. (1993), *Cartography: Thematic Map Design (Third Edition)*, William C. Brown, Dubuque, IA.

Dorling, D. (1995), *A New Social Atlas of Great Britain*, John Wiley and Sons, New York, NY.

Few, S. (2004), *Show Me the Numbers: Designing Tables and Graphs to Enlighten*, Analytics Press, Oakland, CA.

Freedman, D., Pisani, R. & Purves, R. (2007), *Statistics (Fourth Edition)*, W. W. Norton & Company, New York, NY.

Friendly, M. (2000*a*), 'Re–Visions of Minard', *Statistical Computing and Statistical Graphics Newsletter* **11**(1), 1 & 13–19.

Friendly, M. (2000*b*), *Visualizing Categorical Data*, SAS Publishing, Cary, NC.

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.

Gebreab, S. Y., Gillies, R. R., Munger, R. G. & Symanzik, J. (2008), 'Visualization and Interpretation of Birth Defects Data Using Linked Micromap Plots', *Birth Defects Research (Part A): Clinical and Molecular Teratology* **82**, 110–119.

Gentleman, J. F. (1977), 'It's All a Plot (Using Interactive Computer Graphics in Teaching Statistics)', *The American Statistician* **31**(4), 166–175.

Gordon, F. S. & Gordon, S. P. (1992), Sampling + Simulation = Statistical Understanding: Computer Graphics Simulations of Sampling Distributions, *in* F. Gordon & S. Gordon, eds, 'Statistics for the Twenty–First Century (MAA Notes, Number 26)', The Mathematical Association of America, Washington, D.C., pp. 207–216.

Hankins, T. L. (1999), 'Blood, Dirt, and Nomograms: A Particular History of Graphs', *Isis* **90**(1), 50–80.

Harris, R. L. (1999), *Information Graphics — A Comprehensive Illustrated Reference*, Oxford University Press, New York, NY.

Harrower, M. A. & Brewer, C. A. (2003), 'ColorBrewer.org: An Online Tool for Selecting Color Schemes for Maps', *The Cartographic Journal* **40**(1), 27–37.

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.

Hofmann, H. (2007), 'Interview with a Centennial Chart', *Chance* **20**(2), 26–35.

Holcomb, J. & Spalsbury, A. (2005), 'Teaching Students to Use Summary Statistics and Graphics to Clean and Analyze Data', *Journal of Statistics Education* **13**(3), . `http://www.amstat.org/publications/jse/v13n3/datasets.holcomb.html`.

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.

Inselberg, A. (1985), 'The Plane with Parallel Coordinates', *The Visual Computer* **1**, 69–91.

Jones, G. E. (2000), *How to Lie with Charts*, toExcel Press, Lincoln, NE.

Kleiner, B. & Hartigan, J. A. (1981), 'Representing Points in Many Dimensions by Trees and Castles (With Discussion)', *Journal of the American Statistical Association* **76**, 260–276.

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.

Krieger, H. & Pinter-Lucke, J. (1992), Computer Graphics and Simulations in Teaching Statistics, *in* F. Gordon & S. Gordon, eds, 'Statistics for the Twenty–First Century (MAA Notes, Number 26)', The Mathematical Association of America, Washington, D.C., pp. 198–206.

Leisch, F. (2002), Sweave: Dynamic Generation of Statistical Reports Using Literate Data Analysis, *in* W. Härdle & B. Rönz, eds, 'COMPSTAT 2002: Proceedings in Computational Statistics', Physica–Verlag, Heidelberg, pp. 575–580.

Leslie, M. (2002), 'Tools: A Site for Sore Eyes', *Science* **296**(5567), 435.

MacEachren, A. M., Brewer, C. A. & Pickle, L. W. (1995), Mapping Health Statistics: Representing Data Reliability, *in* 'Proceedings of the 17th International Cartographic Conference, Barcelona, Spain, September 3–9, 1995', Institut Cartographic de Catalunya, Barcelona, Spain, pp. 311–319.

MacEachren, A. M., Brewer, C. A. & Pickle, L. W. (1998), 'Visualizing Georeferenced Data: Representing Reliability of Health Statistics', *Environment and Planning A* **30**(9), 1547–1561.

Minnotte, M. C. & West, R. W. (1998), The Data Image: A Tool for Exploring High Dimensional Data Sets, *in* '1998 Proceedings of the Section on Statistical Graphics', American Statistical Association, Alexandria, VA, pp. 25–33.

Monmonier, M. (1993), *Mapping It Out: Expository Cartography for the Humanities and Social Sciences*, University of Chicago Press, Chicago, IL.

Monmonier, M. (1996), *How to Lie with Maps (Second Edition)*, University of Chicago Press, Chicago, IL.

Murrell, P. (2006), *R Graphics*, Chapman & Hall/CRC, Boca Raton, FL.

Olsen, A. R., Carr, D. B., Courbois, J. P. & Pierson, S. M. (1996), Presentation of Data in Linked Attribute and Geographic Space, *in* '1996 Abstracts, Joint Statistical Meetings, Chicago, Illinois', American Statistical Association, Alexandria, VA, p. 271.

Palmer, S. E. (1999), *Vision Science, Photons to Phenomenology*, The MIT Press, Cambridge, MA.

Pickle, L. W. (2008), 'Commentary on "Improving Graphic Displays by Controlling Creativity"', *Chance* **21**(2), 53.

Pickle, L. W. & Herrmann, D. J. (1999), Cognitive Research for the Design of Statistical Rate Maps, *in* '1999 Proceedings of the Section on Survey Research Methods', American Statistical Association, Alexandria, VA, pp. 186–191. `http://www.amstat.org/sections/SRMS/proceedings/papers/1999_029.pdf`.

Pickle, L. W., Mungiole, M., Jones, G. K. & White, A. A. (1996), *Atlas of United States Mortality*, U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, Hyattsville, MD.

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.

Rensink, R. A. (2006), Attention, Consciousness, and Data Display, *in* '2006 JSM Proceedings', American Statistical Association, Alexandria, VA, pp. 2412–2421. (CD).

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.

Robinson, A., Sale, R. & Morrison, J. (1978), *Elements of Cartography (Fourth Edition)*, John Wiley and Sons, New York, NY.

Rosenbaum, A. S., Axelrad, D. A., Woodruff, T. J., Wei, Y.-H., Ligocki, M. P. & Cohen, J. P. (1999), 'National Estimates of Outdoor Air Toxics Concentrations', *Journal of the Air and Waste Management Association* **49**, 1138–1152.

Snell, J. L. & Peterson, W. P. (1992), Does the Computer Help us Understand Statistics?, *in* F. Gordon & S. Gordon, eds, 'Statistics for the Twenty–First Century (MAA Notes, Number 26)', The Mathematical Association of America, Washington, D.C., pp. 167–188.

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. (2006), William Playfair and the Psychology of Graphs, *in* '2006 JSM Proceedings', American Statistical Association, Alexandria, VA, pp. 2426–2436. (CD).

Swayne, D. F., Temple Lang, D., Buja, A. & Cook, D. (2003), 'GGobi: Evolving from XGobi into an Extensible Framework for Interactive Data Visualization', *Computational Statistics & Data Analysis: Special Issue on Data Visualization* **43**(4), 423–444.

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., Axelrad, D. A., Carr, D. B., Wang, J., Wong, D. & Woodruff, T. J. (1999), HAPs, Micromaps and GPL — Visualization of Geographically Referenced Statistical Summaries on the World Wide Web, *in* 'Annual Proceedings (ACSM–WFPS–PLSO–LSAW 1999 Conference CD)', American Congress on Surveying and Mapping.

Symanzik, J. & Carr, D. B. (2008), Interactive Linked Micromap Plots for the Display of Geographically Referenced Statistical Data, *in* C. Chen, W. Härdle & A. Unwin, eds, 'Handbook of Data Visualization', Springer, Berlin, Heidelberg, pp. 267–294 & 2 Color Plates.

Symanzik, J., Carr, D. B., Axelrad, D. A., Wang, J., Wong, D. & Woodruff, T. J. (1999), Interactive Tables and Maps — A Glance at EPA's Cumulative Exposure Project Web Page, *in* '1999 Proceedings of the Section on Statistical Graphics', American Statistical Association, Alexandria, VA, pp. 94–99.

Symanzik, J., Gebreab, S., Gillies, R. & Wilson, J. (2003), Visualizing the Spread of West Nile Virus, *in* '2003 Proceedings', American Statistical Association, Alexandria, VA. (CD).

Symanzik, J., Wong, D., Wang, J., Carr, D. B., Woodruff, T. J. & Axelrad, D. A. (2000), Web–based Access and Visualization of Hazardous Air Pollutants, *in* 'Geographic Information Systems in Public Health: Proceedings of the Third National Conference August 18–20, 1998, San Diego, California', Agency for Toxic Substances and Disease Registry. `http://www.atsdr.cdc.gov/GIS/conference98/`.

Theus, M. (2002), 'Interactive Data Visualization Using Mondrian', *Journal of Statistical Software* **7**(11). `http://www.jstatsoft.org/v07/i11/`.

Theus, M. (2003), 'Abstract: Interactive Data Visualization Using Mondrian', *Journal of Computational and Graphical Statistics* **12**(1), 243–244.

Theus, M. & Urbanek, S. (2004), 'iPlots : Iteractive Graphics for R', *Statistical Computing and Statistical Graphics Newsletter* **15**(1), 11–14.

Theus, M. & Urbanek, S. (2009), *Interactive Graphics for Data Analysis: Principles and Examples*, Chapman & Hall/CRC, Boca Raton, FL.

Tijms, H. (1992), Exploring Probability and Statistics Using Computer Graphics, *in* F. Gordon & S. Gordon, eds, 'Statistics for the Twenty–First Century (MAA Notes, Number 26)', The Mathematical Association of America, Washington, D.C., pp. 189–197.

Tufte, E. R. (1983), *The Visual Display of Quantitative Information*, Graphics Press, Cheshire, CT.

Tufte, E. R. (1990), *Envisioning 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.

Venables, W. N. & Ripley, B. D. (2002), *Modern Applied Statistics with S (Fourth Edition)*, Springer, New York, NY.

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. (2008), 'Improving Graphic Displays by Controlling Creativity', *Chance* **21**(2), 46–52.

Wainer, H. (2009), 'A Centenary Celebration for Will Burtin: A Pioneer of Scientific Visualization', *Chance* **22**(1), 51–55.

Wainer, H. & Francolini, C. M. (1980), 'An Empirical Inquiry Concerning Human Understanding of Two–Variable Color Maps', *The American Statistician* **34**(2), 81–93.

Wallgren, A., Wallgren, B., Persson, R., Jorner, U. & Haaland, J.-A. (1996), *Graphing Statistics & Data: Creating Better Charts*, Sage Publications, Thousand Oaks, CA.

Wang, X., Chen, J. X., Carr, D. B., Bell, B. S. & Pickle, L. W. (2002), 'Geographic Statistics Visualization: Web–based Linked Micromap Plots', *Computing in Science & Engineering* **4**(3), 90–94.

Wegman, E. J. (1990), 'Hyperdimensional Data Analysis Using Parallel Coordinates', *Journal of the American Statistical Association* **85**, 664–675.

Wegman, E. J. & Shen, J. (1993), 'Three–Dimensional Andrews Plots and the Grand Tour', *Computing Science and Statistics* **25**, 284–288.

Yoshioka, K. (2002), 'KyPlot — A User–Oriented Tool for Statistical Data Analysis and Visualization', *Computational Statistics* **17**(3), 425–437.

Zeileis, A., Hornik, K. & Murrell, P. (2008), 'Escaping RGBland: Selecting Colors for Statistical Graphics', *Computational Statistics & Data Analysis* **?**(?), Forthcoming. Preprint available at `http://statmath.wu-wien.ac.at/~zeileis/papers/Zeileis+Hornik+Murrell-2008.pdf`.

Zelazny, G. (2001), *Say it with Charts: The Executive's Guide to Visual Communication (Fourth Edition)*, McGraw–Hill, New York, NY.

— THE END —

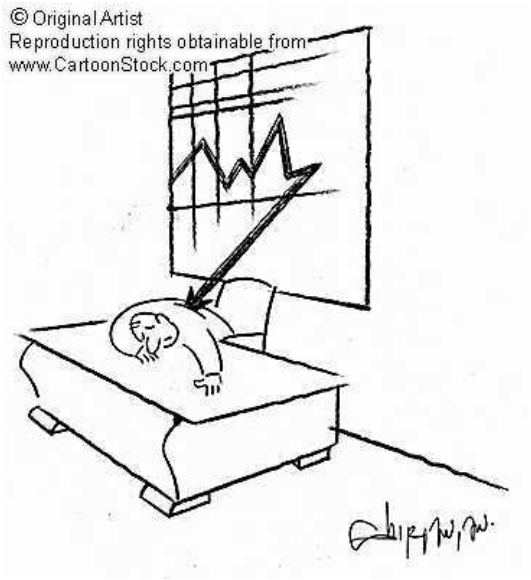


Figure 102: `http://www.cartoonstock.com/blowup_stock.asp?imageref=vsh0184&artist=Shirvanian,+Vahan&topic=statistics+`, Cartoon.