Visualizing and Exploring Data

Based on Chapter 3 of Hand, Manilla, & Smyth

David Madigan
Introduction

• Exploratory Data Analysis legitimized by Tukey (1997)
• ~ Data based hypothesis generation
• Always need to be skeptical about findings since the search space can be very large
• Useful tools: S-Plus, Ggobi, DataDesk, JMP

http://otal.umd.edu/Olive/

(On-line Library of Information Visualization Environments)
Displaying Single Variables

credit card example
Pima indian example
Smoothing Estimates

• Kernel estimates smooth out the contribution of each datapoint over a local neighborhood of that point.

\[ \hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-x(i)}{h}\right) \]

\( h \) is the kernel width

• Gaussian kernel is common:

\[ Ce^{-\frac{1}{2}\left(\frac{x-x(i)}{h}\right)^2} \]

• Formal procedures for optimal bandwidth choice
• Gray & Moore’s work on speeding this up…
Displaying Two Variables
Figure 3.7: A scatterplot of 96,000 cases, with much overprinting. Each data point represents an individual applicant for a loan. The vertical axis shows the age of the applicant, and the horizontal axis indicates the day on which the application was made.
Correlation = 0.5
Corridor Effect: None (N), Main (M), Interaction (I), Both (B), Error in Feet

N=5  N=10  N=20  N=50  N=100  N=253

BR Data  BR Data  BR Data  BR Data  BR Data  BR Data

0  10  20  30  40  50

N  I

N=5  N=10  N=20  N=50  N=145

CA Down Data  CA Down Data  CA Down Data  CA Down Data  CA Down Data

0  50  100  150  200

N  I

N=5  N=10  N=20  N=55

CA Up Data  CA Up Data  CA Up Data  CA Up Data

0  50  100  150

N  I
Figure 4: Small Scale Simulation Results. Each boxplot shows a particular setting of $\theta$, $d_F$, and $d_S$. The horizontal axes show the log-ratio of the mean square error from random sampling to the mean square error from LDS.
Tinting

• Experiment to model the effects of car window tinting on visual performance

• csoa: critical stimulus onset asynchrony (time to recognize an alphanumeric target)

• it: inspection time (time required for a simple discrimination task)

• age, tint (no,lo,hi), target (locon,hicon), sex
xyplot(csoa ~ it | sex * agegp, data = tinting, groups = target, auto.key = list(columns = 2))
xyplot(csoa~it | sex*agegp, data=tinting, groups=tint, auto.key=list(columns=3))
xyplot(csoa~it | sex*agegp, data=tinting, groups=tint, auto.key=list(columns=3), type=c("p","smooth"), span=0.8)
Causes of Mortality in the Army in the East
April, 1854 to March 1855

From: F. Nightingale, "Notes on Matters Affecting the Health, Efficiency and Hospital Administration of the British Army", 1858

Source: Michael Friendly
Half-space location depth of $\tilde{z}$ in $\mathbb{R}^2$ relative to $z_1, \ldots, z_n$ is the smallest number of $z_i$ contained in any closed half-plane with boundary line through $\tilde{z}$.
Figure 1: Fourfold display for Berkeley admissions data: Evidence for sex bias? The area of each shaded quadrant shows the frequency, standardized to equate the margins for sex and admission. Circular arcs show the limits of a 99% confidence interval for the odds ratio.
Four-fold display for categorical data

Figure 2: Fourfold display of Berkeley admissions, by department. In each panel the confidence rings for adjacent quadrants overlap if the odds ratio for admission and sex does not differ significantly from 1. The data in each panel have been standardized as in Figure 1.
Mosaic plots for categorical data

Table 1: Hair-color eye-color data

<table>
<thead>
<tr>
<th>Eye</th>
<th>Hair Color</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Black</td>
<td>Brown</td>
<td>Red</td>
<td>Blond</td>
<td>Total</td>
</tr>
<tr>
<td>Green</td>
<td>5</td>
<td>29</td>
<td>14</td>
<td>16</td>
<td>64</td>
</tr>
<tr>
<td>Hazel</td>
<td>15</td>
<td>54</td>
<td>14</td>
<td>10</td>
<td>93</td>
</tr>
<tr>
<td>Blue</td>
<td>20</td>
<td>84</td>
<td>17</td>
<td>94</td>
<td>215</td>
</tr>
<tr>
<td>Brown</td>
<td>68</td>
<td>119</td>
<td>26</td>
<td>7</td>
<td>220</td>
</tr>
<tr>
<td>Total</td>
<td>108</td>
<td>286</td>
<td>71</td>
<td>127</td>
<td>592</td>
</tr>
</tbody>
</table>

Diagram illustrating mosaic plot for the table data.
Visual Scalability
Eick and Karr

• Human perception: 6.5 million pixels?

• Monitor resolution: 640X480 = 307,300; 1600X1200 = 1,920,000

• Visual metaphors:
  • Bar charts: can display 500; realistic limit about 50; color
  • Matrix views: 1280X1024 can display 13,000 10X10 entities
  • Landscapes: 3-D matrix view; color, height, and shape; occlusion?
  • Network views: scalability depends on connectivity
  • Scatterplots: 100,000 points?
  • Histograms: smoothing calculations become expensive

• Interactivity
Dimensionality Reduction

• Scatterplot = 2-D projection defined by 2 variables (e.g. \(x_1\) vs. \(x_4\))

• Other projections? e.g. \(2x_1 + 3x_2\) vs. \(6x_1 + 2x_4\)

• Projection pursuit…issues with scalability

• PCA: scales quite well; used in text retrieval

• MDS: metric versus non-metric

• Random projections
Tufte:

Graphical excellence is the well-designed presentation of interesting data - a matter of substance, of statistics, and of design.

Graphical excellence consists of complex ideas communicated with clarity, precision, and efficiency.

Graphical excellence is that which gives the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space.

Graphical excellence is nearly always multivariate.

And graphical excellence requires telling the truth about the data.
Tufte also insists that graphical displays should:

induce the viewer to think about the substance rather than about methodology, graphic design, the technology of graphic production or something else

reveal the data at several levels of detail, from a broad overview to the fine structure
CARTE FIGURATIVE des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.

Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.

TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.

Xh. = December  
G. = November  
G. = October
In the following example, from The Times of Saturday 1/2/3 is a superb example of this form of abuse. The two shells supposedly represent two quantities in the ratio 500 to 364, so the first should be 500/364 or 1.374 times bigger than the second, representing a 37.4% increase. But their lengths are in the ratio 102mm to 65mm, making the first 1.569 times longer than the second, and giving it a volume greater than that of the second by a factor of 1.569 cubed, or 3.864. This gives a shocking lie factor of 3.864/1.374 or 2.8 times!
This pie chart shows time spent with doctors. Use it to answer questions 4 to 7.

4) Which two countries give their patients the most time?
5) Which two countries give their patients the least time?
6) What colour is the UK slice?
7) Which country gives their patients about the same amount of time as the UK?

Now check your answers with those on the answer sheet.
Tufte’s worst graphic ever!
Edward R. Tufte

The Cognitive Style of PowerPoint:
Pitching Out Corrupts Within

Military parade, Stalin Square, Budapest, April 4, 1956.
DESTRUCTION OF ONE OF THE GREATEST MODERN CITIES.

BIRD’S-EYE VIEW OF STRICKEN SAN FRANCISCO, SHOWING THE BURNED DISTRICT, COVERING TWENTY-FIVE SQUARE MILES, WITH THE MOST PROMINENT PLACES AND BUILDINGS CAREFULLY INDICATED.—Drawn especially for Leslie’s Weekly by H. M. Pitts.
http://www.ted.com/talks/view/id/92